# Data-driven quality improvement approach to reducing waste in manufacturing

Rose Clancy
*Civil Engineering, University College Cork, Cork, Ireland, and*
Dominic O'Sullivan and Ken Bruton
*School of Engineering, University College Cork, Cork, Ireland*

Received 26 February 2021
Revised 30 April 2021
16 June 2021
Accepted 14 July 2021

## Abstract

**Purpose** – Data-driven quality management systems, brought about by the implementation of digitisation and digital technologies, is an integral part of improving supply chain management performance. The purpose of this study is to determine a methodology to aid the implementation of digital technologies and digitisation of the supply chain to enable data-driven quality management and the reduction of waste from manufacturing processes.

**Design/methodology/approach** – Methodologies from both the quality management and data science disciplines were implemented together to test their effectiveness in digitalising a manufacturing process to improve supply chain management performance. The hybrid digitisation approach to process improvement (HyDAPI) methodology was developed using findings from the industrial use case.

**Findings** – Upon assessment of the existing methodologies, Six Sigma and CRISP-DM were found to be the most suitable process improvement and data mining methodologies, respectively. The case study revealed gaps in the implementation of both the Six Sigma and CRISP-DM methodologies in relation to digitisation of the manufacturing process.

**Practical implications** – Valuable practical learnings borne out of the implementation of these methodologies were used to develop the HyDAPI methodology. This methodology offers a pragmatic step by step approach for industrial practitioners to digitally transform their traditional manufacturing processes to enable data-driven quality management and improved supply chain management performance.

**Originality/value** – This study proposes the HyDAPI methodology that utilises key elements of the Six Sigma DMAIC and the CRISP-DM methodologies along with additions proposed by the author, to aid with the digitisation of manufacturing processes leading to data-driven quality management of operations within the supply chain.

**Keywords** Digitisation, Digital manufacturing, Six sigma, CRISP-DM, Quality improvement, Data mining

**Paper type** Research paper

## 1. Introduction

There has been a major shift towards automation and digitisation in the manufacturing industry. The automation and digitisation of all business processes is a fundamental part of Industry 4.0 (Telukdarie *et al.*, 2018). A systematic mapping study of the literature in the area of digitisation and analysis of manufacturing processes found that over half of the papers in the area stated that transforming to digital manufacturing improves the efficiency of manufacturing processes (Clancy *et al.*, 2020). Research also shows that the majority of

This publication has emanated from research that was part-funded by Science Foundation Ireland (SFI) through MaREI, the SFI Research Centre for Energy, Climate, and Marine [Grant No: 12/RC/2302_P2], with supporting funding obtained from DePuy Synthes. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Emerald Publishing Limited
1754-2731
DOI 10.1108/TQM-02-2021-0061

productivity increases in organisations today, originate directly or indirectly, from digitisation and big data analytics (Weill and Woerner, 2015). Other benefits of digitisation in manufacturing are, the ability to have real-time data monitoring, a reduction of quality costs and improved product quality (Clancy *et al.*, 2020). Digitisation helps to reduce uncertainties through the increased visibility and transparency of the supply chain (Bag *et al.*, 2020c). In the era of digital transformation, the effective management of data enables organisations to leverage the power of big data analytics to deliver high quality products and services and effectively manage their supply chain operations (Bag *et al.*, 2020b). The use of big data in business analytics to inform decisions in operations management is a key benefit of digitising manufacturing processes (Bag *et al.*, 2020b). The term "big data" refers to a very large volume of data, and big data analytics is the management of such data (Dhamija and Bag, 2020). Manufacturing operations have a multitude of unstructured data from various sources. The digitisation of this data, along with efficiently functioning information technology, enables business analysts to extract key insights from the supply chain operations and aid managers to make data-driven decisions, hence improving supply chain management (Dhamija and Bag, 2020; Bag *et al.*, 2020a).

Although this trend of digitisation and big data analytics has gained a lot of attention, it is still not clear how it can actually be implemented in supply chain operations management (Petersen *et al.*, 2017). Furthermore, big data analytics in supply chain management presents a relatively new research area, and there is a lack of theoretical studies on the application of big data analytics in quality management (Akter *et al.*, 2016). To improve supply chain management, and achieve data-driven quality management, real-time data from manufacturing processes is required to provide precise predictions of product quality (Belhadi *et al.*, 2019). In summary, the research gaps are as follows. First, it is not clear how to implement digitisation in an interoperable way (Petersen *et al.*, 2017). Second, manufacturing digitisation in the context of Industry 4.0 is a novel research direction (Johansson *et al.*, 2019). Lastly, there is a lack of theoretical studies on the use of big data analytics to achieve data-driven quality management for improving supply chain management (Akter *et al.*, 2016). Furthermore, Elg *et al.* discusses the problems and prospects of digitalisation and quality management and the heavy reliance on the need for IT skills (Elg *et al.*, 2020). Without digitisation, process experts will continue to carry out intensely time-consuming manual ad-hoc analysis. It is not possible to analyse all the combinations of data tags continuously, manually from a production line; results are often not shared across the organisation, and this type of analysis is not sufficient for manufacturing firms to remain competitive in the era of Industry 4.0. Therefore, the purpose of this study is to determine an approach that managers can utilise to digitise their operations, enabling data-driven quality management of their manufacturing operations in the supply chain. Existing quality management concepts do not cater for the digitisation of manufacturing processes in terms of skills, resources and the method or approach required. To address this objective, the research question for this study is:

*RQ1*. How can industrial practitioners use existing quality management concepts in the digitisation of their supply chain operations to achieve data-driven quality management and improved supply chain performance?

This study will therefore provide a digitisation methodology for managers to implement in their supply chain operations to achieve improved operations management and quality management through data-driven decision-making. The remainder of the paper is organised as follows. The next section provides the literature review of traditional manufacturing improvement methods. Section 3 consists of a review and comparison of existing methods in relation to quality management and data mining. Section 4 briefly outlines the problems faced when trying to implement existing quality management methods to digitise and manage supply chain operations using data-driven decision-making. Section 5 presents the proposed methodology for the implementation of digitisation and data-driven quality management of supply chain operations. Section 6 presents the outcomes and achieved objectives, the managerial implications, future research proposals and limitations of the study.

## 2. Literature review

There is a problem with "traditional" methods of analysis for industrial manufacturing processes. An example is root cause analysis which is the process of identifying factors that cause defects or deviations in quality of the manufactured product (Ge *et al.*, 2017). The nature of manufacturing data is dynamic and complex, and because of this, it simply is not feasible for process experts to keep track of all of this data (Ge *et al.*, 2017). In order to perform efficient, scalable root cause analysis for manufacturing processes in industry, the power of machine-learning models must be infused with process expertise (Ge *et al.*, 2017). Data are the building block upon which organisations thrive, and it is essential that the true data relating to the process are used to drive decision-making (Chong and Shi, 2015). Data analytics is used to extract useful and hidden patterns and important relationships from large data sets that were previously unknown (Chong and Shi, 2015). With the ever-increasing amount of data generated in organisations, there is a greater need for efficient and effective ways of analysing data. Having vast amounts of data are not enough to make data-driven decisions, as these data sets can no longer be easily analysed. There is now a need for new tools and methods for analysing big data in organisations (Chong and Shi, 2015). In many organisations, data are not considered because the analysis is too time-consuming, expensive or there is a lack of understanding of how to analyse the data to find valuable, actionable insights. Many managers in organisations actually lack confidence that analytics will improve their decision-making (Court, 2015). This can be because they do not understand the analytics or the recommendations it suggests (Court, 2015). When managers do not trust the analytics they fall back on the historic rule of thumb that has always been used (Court, 2015). One of the reasons that organisations are not performing true data-driven decision-making for their manufacturing processes is because the customer, e.g. the engineers of the manufacturing process, is not involved in the analysis. In many cases, the company outsources the analysis or directs it to the internal analytics team, this results in the management pushing back on the findings of the analysis either because it is too complex or because there is a lack of transparency in the analysis process (Court, 2015). Data mining and analytics play a major role in the decision-making for quality management of manufacturing processes (Ge *et al.*, 2017), however it is crucial that the human-data intelligence of the process experts is utilised. Human-data intelligence refers to the experience and knowledge relating to the manufacturing process, and it is a 'must have' for developing big data analytics capabilities (Bang *et al.*, 2019). To overcome this problem, a data-driven quality improvement methodology is needed to digitise key information relating to the process and combine the skills from data science along with the knowledge of the manufacturing process. Existing methodologies in relation to quality improvement of a supply chain and data science are investigated in the following section.

### *2.1 Traditional process improvement methods*

Over the last number of decades, different quality management concepts have been applied in the manufacturing industry. One of the definitions of quality management is "a comprehensive way to improve total organisation performance" (Foley, 2004). Quality management concepts include TQM, Six Sigma, lean manufacturing, business process re-engineering (BPR), just-in-time, Kaizen and business excellence (Andersson *et al.*, 2006).

Although these quality management concepts differ, they all share a similar purpose, that is, to improve processes by minimising waste while improving product quality and reducing quality costs. This study focuses on the area of quality improvement within quality management, as it has been significantly affected by the new era of digitalisation (Elg *et al.*, 2020). Digital technology offers great potential in achieving internal process improvements (Elg *et al.*, 2020), and it is vital for companies to utilise the potential of already available data to proactively control their processes. Only those companies that can analyse their business operations based on the rapidly growing quantity of data and predict the optimal process conditions will survive in the highly competitive manufacturing environment (Krumeich *et al.*, 2014). The aim of data mining in manufacturing is to obtain useful information from process data and convert it to effective knowledge for decision-making leading to process improvement (Ge *et al.*, 2017). There have been multiple studies focused on advancing existing quality improvement methods to be suitable for the new era of digital manufacturing. Mayr *et al.* (2018) discusses combining lean management and Industry 4.0, leading to a connected manufacturing environment where data can be transmitted in real-time. Schafer *et al.* (2018) developed a quality management-CRISP-DM cycle with a communication aspect in every phase of the cycle to overcome domain barriers. These studies are certainly valuable and highlight the need to renovate the existing quality management concepts used for process improvement and exploit the additional benefits of digital transformation. Big data analytics has been identified as playing a critical role of Industry 4.0 (Sharma and Pandey, 2019), and it has the potential to contribute to improved performance of the supply chain (Wagner *et al.*, 2017). An example of one of the benefits brought about by digital transformation is the ability to analyse real-time process information to gain insights into process parameters and identify trends (Mayr *et al.*, 2018). While previous studies by Mayr *et al.* (2018) and Schafer *et al.* (2018) have looked at combining quality management concepts and data mining methods, they do not focus on the digitisation of the process, which is necessary to enable real-time analysis and optimisation of processes using data-driven decision-making. In the next section, conventional quality improvement concepts and data mining methods are assessed for suitability in terms of implementing digitisation in manufacturing with data-driven decision-making to increase the efficiency of processes and reduce waste. A decision matrix, which is a decision-making tool that evaluates a list of options based on several criteria, is used to assess quality improvement concepts and data mining methods. The three well-known quality improvement concepts, TQM, Six Sigma and lean manufacturing are assessed. TQM has been described by Dahlgaard *et al.* (1997) as a corporate culture characterised by increased customer satisfaction through continuous improvement. TQM has also been described by Klefsjö *et al.* (2001) as a continuously evolving management system consisting of values, methodologies and tools with the aim being to increase customer satisfaction with a reduced amount of resources. Klefsjö *et al.* (2001) define methodologies as "ways to work within the organisation to reach the values". For simplicity, when assessing the different concepts, methods and processes for quality improvement and data mining, they will all be referred to as methodologies in this paper. The objective of the second quality improvement methodology, lean manufacturing, is to reduce waste by improving process flow. The third quality improvement methodology is Six Sigma. It is a problem-solving methodology designed to improve quality and reduce variation in a process. It started as a quality measurement approach which developed into a process improvement methodology, and it is now commonly seen as the philosophy of many organisations (Vinet and Zhedanov, 2011). A second decision matrix is used to assess three data mining methodologies. The three data mining methodologies assessed are the three most common data mining methodologies used by large enterprises (Shafique and Qaiser, 2014). They are the cross-industry standard process for data mining (CRISP-DM), sample, explore, modify, model and assess (SEMMA) and knowledge discovery databases (KDD). CRISP-DM is a data science methodology for designing, creating and building, testing and deploying machine-learning solutions. It provides a structured approach to planning a data mining project (Chapman *et al.*, 2000). SEMMA is a list of steps developed by Statistical Analysis System (SAS) to guide the implementation of a data mining project. Lastly, KDD is a broad process of finding knowledge in data, and it focuses on the high-level application of data mining methods (Fayyad *et al.*, 1996).

## 3. Methodology

### *3.1 Review of existing methods*

The criteria chosen to assess the quality improvement and data mining methodologies were, first, how widely used the methodology is, secondly, how applicable it is to the objective of the study and lastly, how adaptable it is. The objective of the study is to determine a concrete approach to digitising a manufacturing process to improve and control processes and achieve data-driven quality improvement. This study aims to integrate quality improvement and data mining methodologies to digitise and improve manufacturing processes using data-driven decision-making. The third criterion, adaptability, was included in the assessment of the methodologies because the process of analytics is not linear, and it is an iterative cycle in which the answers to the initial questions almost always generate additional questions (McCue, 2015). Therefore, it is important that the methodology is adaptable, i.e. an iterative process. Both the quality improvement and data mining methodologies were scored from 1–3 (1 = poor, 2 = average, 3 = good) for each of the criteria as seen in Tables 1 and 2. Six Sigma and lean are both widely used methodologies in comparison to TQM. Efforts to implement TQM have been unsuccessful in many organisations (Eskildson, 1994; Foley, 2004); therefore, TQM scored the lowest of the quality improvement concepts for widespread use. Whereas, Six Sigma is now established in almost every industry and lean manufacturing is a wide-spread and successful concept (Andersson *et al.*, 2006), giving them both a score of 3. The purpose of this study is to combine the traditional quality improvement concepts with data mining to aid organisations in transitioning to Industry 4.0, specifically, using digitisation and data-driven decision-making to improve manufacturing processes. Six Sigma is a data-driven improvement cycle used for optimising business processes and reducing the number of defects (Knowles, 2011), and the combination of Six Sigma with data mining tools could enhance the existing limited statistical experimental design as part of Six Sigma (Schafer *et al.*, 2018). Schafer *et al.* (2019) suggested that the combination of Six Sigma and data mining methods could help to move analytics from being purely descriptive to predictive. Therefore, it scored 3 for the applicability criterion. The integration of lean manufacturing with Industry

| Methodology | Widely used | Applicability | Adaptability | Total |
|---|---|---|---|---|
| Six sigma | 3 | 3 | 3 | 9 |
| TQM | 1 | 1 | 3 | 5 |
| Lean | 3 | 2 | 1 | 6 |

**Table 1.** Decision matrix for quality management methodologies

| Methodology | Widely used | Applicability | Adaptability | Total |
|---|---|---|---|---|
| CRISP-DM | 3 | 3 | 3 | 9 |
| SEMMA | 2 | 1 | 3 | 7 |
| KDD | 2 | 2 | 3 | 4 |

**Table 2.** Decision matrix for data mining methodologies

4.0 has had mixed reviews in the literature (Buer *et al.*, 2018). Some authors have said that that lean manufacturing is a prerequisite to the successful introduction of Industry 4.0 (Huber, 2016), Mrugalska and Wyrwicka (2017) believe that Industry 4.0 and lean can coexist and support each other, and Rüttimann and Stöckli (2016) believe that Industry 4.0 will improve the flexibility of lean production systems. As confusion remains in how lean manufacturing can be combined with data mining, it scored more than TQM, but less than Six Sigma for the applicability to this study. Six Sigma is an iterative process (Knowles, 2011); therefore, it scored the highest score of 3 for the adaptability criterion. A fundamental principle of TQM relates to continuous improvement of the product (Middleton, 1996), therefore, this quality improvement methodology also scored 3 for the adaptability criterion. Lean is less flexible and iterative in nature, and it cannot deal with highly dynamic conditions (Andersson *et al.*, 2006); therefore, it scored 1 for the adaptability criterion. In summary, as Six Sigma scored the highest in the decision matrix, it was selected as the most suitable quality improvement methodology for this study. In the assessment of the three data mining methodologies, CRISP-DM scored the highest for the first criterion, as it is the most commonly used methodology in data science (Kristoffersen *et al.*, 2019) and the most widely applied commercially (Shafique and Qaiser, 2014). The CRISP-DM process includes business understanding and deployment stages that make the process suitable for real-world industrial projects (Pyvovar *et al.*, 2019). The business understanding phase is crucial for manufacturing improvement, as it allows a greater level of understanding of the use case in comparison to SEMMA (Palacios *et al.*, 2017). KDD also requires a prior understanding of the application domain as well as incorporating this knowledge in the system (Fayyad *et al.*, 1996). In conclusion, CRISP-DM is a more complete methodology than SEMMA (Shafique and Qaiser, 2014), and KDD is not necessarily considered as a detailed methodology with steps to follow but an overall high-level process that incorporates data mining (Azevedo and Santos, 2008). Therefore, as CRISP-DM is the most applicable data mining methodology for this study, it scored the highest for this criterion, and since SEMMA is not as complete as CRISP-DM and does not incorporate business understanding, it scored the lowest for the applicability criterion. The phases in the CRISP-DM methodology are iterative and reversible (Palacios *et al.*, 2017), SEMMA is too an iterative methodology as its internal steps can be performed iteratively (Palacios *et al.*, 2017), and KDD is also an iterative and interactive methodology (Shafique and Qaiser, 2014). When working with manufacturing data, unforeseen gaps and errors often arise. This seems to be well accepted in the data science world, in comparison to the manufacturing process improvement domain, as each of the data mining methodologies are iterative in their approach. This helps when working with real data as any errors can be resolved without having to complete the entire cycle (Pyvovar *et al.*, 2019). Hence, each of the data mining methodologies scored 3 for the adaptability criterion. As CRISP-DM scored the highest in the decision matrix, it was selected as the most suitable data mining methodology for this study. In conclusion, Six Sigma and CRISP-DM are the most suitable quality improvement and data mining methodologies for the implementation of digitisation of manufacturing processes with data-driven decision-making leading to process improvements.

### *3.2 Six sigma and CRISP-DM*

The goal of this study is to document a formal approach to digitising a manufacturing process to improve and control processes and achieve data-driven quality improvement of supply chain performance. The highest scoring quality improvement and data mining methodologies from the decision matrix were Six Sigma and CRISP-DM. These two methodologies will be implemented together in an industrial use case to test their effectiveness of digitising the process and using data-driven decision-making to improve the process. The objectives of Six Sigma are to improve quality and reduce variation within a process (Vinet and Zhedanov, 2011). Variation is known as the enemy in a process as it fuels waste and increases cost. To reduce variation in a process, several factors, essentially the inputs to the process need to be controlled. The two types of variation in a process are common cause variation and special cause variation. Common cause variation is random and inherent within your process, whereas special cause variation is explainable and controllable. A key difference in these types of variation is that adjusting the process causes common cause variation to increase and special cause variation to decrease (Vinet and Zhedanov, 2011). Define-measure-analyse-improve-control (DMAIC) is a structured problem-solving process used in Six Sigma projects. Figure 1 illustrates the steps in the DMAIC process. The DMAIC method within Six Sigma is regularly referred to as an approach for solving existing problems, particularly in the manufacturing sector (De Mast and Lokkerbol, 2012). The tools and techniques encompassed within the DMAIC process facilitate organisations to improve their processes using an interdisciplinary approach (Schroeder *et al.*, 2008). Therefore, as this study is focused on manufacturing operations, and combining existing quality improvement methodologies from the engineering discipline with data mining methodologies from the data science discipline, Six Sigma is an appropriate choice. It should be noted, however, that Six Sigma has been challenged by many authors, and it indeed has drawbacks. One of these drawbacks is that the focus of Six Sigma projects is generally for existing customers or processes. Therefore, it might not be useful for fast-paced, highly innovative environments (Basios and Loucopoulos, 2017). However, using the Six Sigma DMAIC approach has shown to generate valuable process improvements in manufacturing organisations (Swarnakar and Vinodh, 2016). As Six Sigma is a general approach, researchers have found the need to enhance the DMAIC method and consider requirements from a multi-disciplinary viewpoint, to address challenges arising from new technology. One example of this is the Six Sigma DMAIC enhanced with capability modelling approach that further presents the need to consider requirements for business, information technology, data science and operations disciplines to bring process improvements and implement digital technology (Basios and Loucopoulos, 2017). Therefore, the Six Sigma DMAIC process will be implemented in an industrial use case in combination with a data mining method, to aid the digitisation and data-driven quality management of a supply chain manufacturing operation. The DMAIC tool will

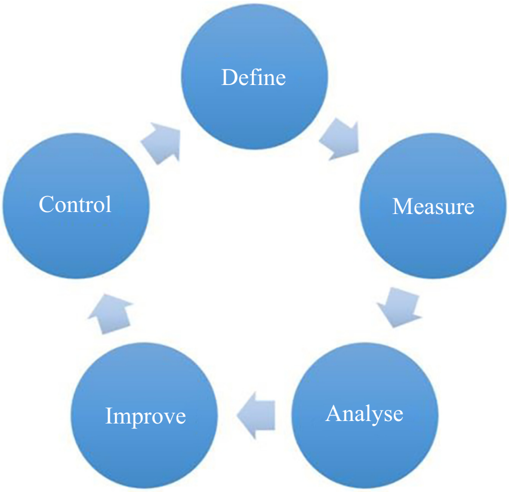


**Figure 1.**
Six Sigma DMAIC process

be used in the industrial use case with the aim of reducing the special cause variability by controlling the inputs to the process. However, before the inputs can be controlled to reduce the variation in the process, the information relating to the process must first be digitised. After digitising the process, the data can be analysed to try to understand the relationships between the variables. For effective data mining and analytics it is vital that relationships and process causality among different elements in the process are identified (Ge *et al.*, 2017). The CRISP-DM approach to data mining projects (seen in Figure 2) will be implemented in the industrial use case to aid in the digitisation of the manufacturing and use insights derived from the analysis to reduce the number of defective products. The CRISP-DM approach comprises of six phases, business understanding, data understanding, data preparation, modelling, evaluation and deployment. Although these phases are performed consecutively, the process is iterative in nature because as knowledge is gained from some of the phases, further iterations can be improved.

## 4. Demonstration in industry

The manufacturing organisation in this use case is in Ireland. It is one of the world's largest manufacturers of orthopaedic implants. They manufacture orthopaedic products for joint replacement, trauma, spine, sports medicine and others. The value stream in this use case is a foundry, which employs investment casting (also known as precision casting) methods to manufacture components of biomedical joint replacements. A foundry is a factory in which castings are produced by melting metal, pouring liquid metal into a mould and then allowing it to solidify. A value stream is all of the steps (both value adding and non-value adding) in a process that are essential to producing a product or service (Vinet and Zhedanov, 2011). The value stream that was focused on in this use case had scrap rates that periodically exceeded

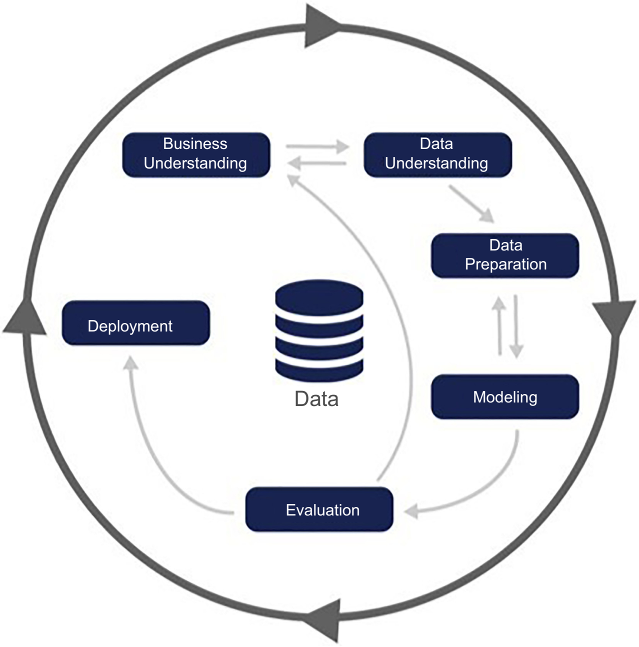


**Figure 2.**
Phases of the CRISP-DM reference model (Chapman *et al.*, 2000)

the expected performance for the investment casting process. The value stream in this organisation has a high focus on increasing their yields and has found this key performance indicator to be a significant challenge. There exists a plethora of potential insights to be gained through analysis of existing data relating to the manufacturing process; however, much of the data are not easily accessible for analysis or even in digital format. This identified the need for increased digitisation and connection of data for this manufacturing process. This in-turn would underpin future problem-solving and continuous improvement activities to further reduce scrap rates. The digitisation of this manufacturing process would enable data-driven quality management and improved supply chain management performance.

### *4.1 Implementation of DMAIC and CRISP-DM*

The Six Sigma and CRISP-DM methodologies were implemented in the industrial use case with the aim of digitising the process and combining the data mining and process improvement methodologies to discover actionable insights that can reduce the high-level of defective products produced. The first stage of the Six Sigma methodology is the define phase and the first step in the CRISP-DM methodology is business understanding. Therefore, the first step in the project was to understand the business problem. It is vital that the problem definition is based on real data and not just opinions. Using a business analytics platform that utilises operating management system data, the business problem was illustrated clearly in graphical format by looking at the weekly percentage of scrap (defective parts) for the value stream. After defining the business problem, a project charter was documented. The project charter outlined the project lead and the objectives of the project. The objective of this project was to perform analysis with the aim of reducing the scrap rate of the value stream. The next phase of the DMAIC process is the measure phase and the next steps in the CRISP-DM methodology are data understanding and data preparation. A process map seen in Figure 3 was created for the manufacturing process containing detailed information for the product types and their sequence of steps in the manufacturing process. The process map illustrates the information that is collected throughout the manufacturing process and the quality specifications that exist such as environmental conditions. Creating the process map helped to achieve the next step from CRISP-DM, data understanding. Manufacturing processes are often quite complex; therefore, it is crucial to understand the detail of the process. The purpose of the measure phase in Six Sigma is to gather information relating to the process. Some initial data was collected which found that for most of the variables related to the value stream, the data were not in a condition suitable for analysis. Some of the variables of the manufacturing process were not digitised, i.e. still manually recorded or paper based. One example of an information source relating to the manufacturing process in this use case that needed to be digitised was the daily material testing data that is recorded on paper. This highlighted a gap in using the two methodologies, as neither the DMAIC nor CRISP-DM methodology gives guidance of how to digitise a manufacturing process. Digitising the process is necessary, as the ability to deliver a solution is dependent on the availability of the necessary data. Therefore, a digitisation plan is needed to allow more information relating to the manufacturing process to be analysed. However, to pinpoint which sources required digitisation, an overall schematic diagram is needed. A data architecture diagram would be beneficial in outlining the various data sources connected to the manufacturing process. The next phase of the DMAIC process is the analyse phase, and the next step in the CRISP-DM process is data preparation. Before conducting the analysis, some data preparation was needed. The data preparation involves formatting and re-structuring of the data so that it can be used for analysis. The data preparation was completed using the spreadsheet, Microsoft Excel. Using this platform for data preparation presented the following issues. Due to the volume of formulae used in the spreadsheet to re-organise and format the data, the

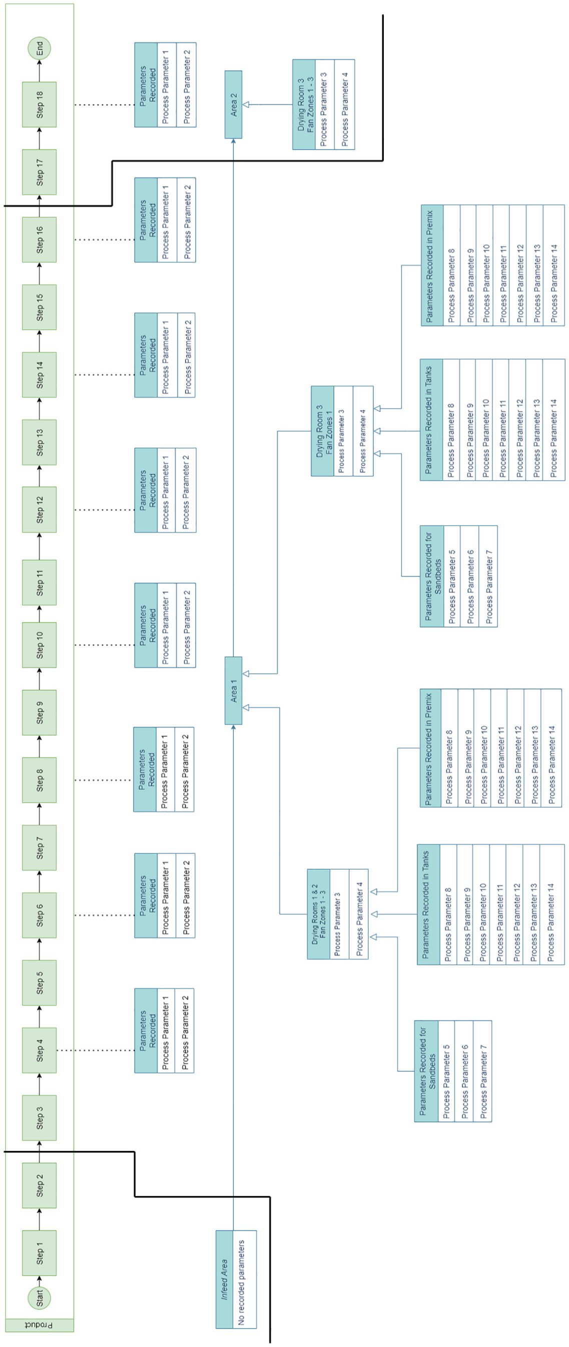


**Figure 3.**
Process map

spreadsheet would stop functioning and abort unexpectedly. Along with this if an error were found in the data, for example the products had been grouped incorrectly, this would require significant time and effort to correct the spreadsheet. Furthermore, it requires an extensive knowledge of the intricacies of the process to detect errors in the data (e.g. common data entry errors). These types of errors can only be recognised if those preparing the data understand these details relating to the process. After preparing the data, some exploratory analysis was completed with the limited data. This brought about an array of questions relating to the process, which resulted in back-and-forth communications with the process experts and multiple iterations and corrections of the analysis. The following are some examples. The time difference between the material test results and the point at which the product was identified as not meeting the specification had to be accounted for in the analysis. Each product family had to be analysed separately. Firstly, due to the variation in process steps across product families and secondly, because the type of scrap reasons for each product family can vary. There are certain chambers within equipment assets in the process used in alteration, and which batches were used in which chamber was not typically recorded, so data relating to the potential influence of each chamber could not be analysed. These are just some of the examples and ultimately, without these pieces of process knowledge, the analysis conducted would not have been accurate. Conducting an in-depth process overview and ensuring the analysts have a thorough understanding of the process before analysis is initiated would be beneficial in conducting efficient and accurate analysis of the process. There were multiple learnings from the implementation of the DMAIC and CRISP-DM methodologies in this use case; although the combination of the two methodologies from the process improvement and data mining disciplines is a huge improvement in helping organisations to transition to digital manufacturing, there remains some areas for improvement. To fill these gaps, the author has proposed several additions to these two methodologies. The proposed methodology is discussed in the next section.

## 5. Proposed HyDAPI methodology

The author proposes the hybrid digitisation approach to process improvement (HyDAPI) methodology in Figure 4. The purpose of the HyDAPI methodology is to guide the digitisation of manufacturing processes to enable data-driven decision-making resulting in reduced waste. The methodology includes specific tasks recommended by the author along with the DMAIC and CRISP-DM methodologies to combat the problems that arose in the industrial use case. There are five phases in this methodology, define, measure, analyse, improve and control originate from the Six Sigma DMAIC process. The concept of having tasks and outputs originates form the CRISP-DM methodology.

### *5.1 Define phase*

The define phase focuses on the problem at hand and the end goal. The first task, **business understanding** focuses on understanding the objectives of the project and the business requirements (Chapman *et al.*, 2000). The second task in the define phase is **collaboration**. Many projects fail due to a lack of communication and collaboration, the specific problem in this study is that a gap was found to exist in industry between the engineers of a process and the data scientists performing the analysis. To successfully achieve the objective of improving the manufacturing process, the author has proposed **collaboration** as a task in the define phase. It is vital that there is a strong collaboration from the beginning of the project between those with knowledge and understanding of the manufacturing process and the data scientists performing the statistical analysis. The project charter and project plan

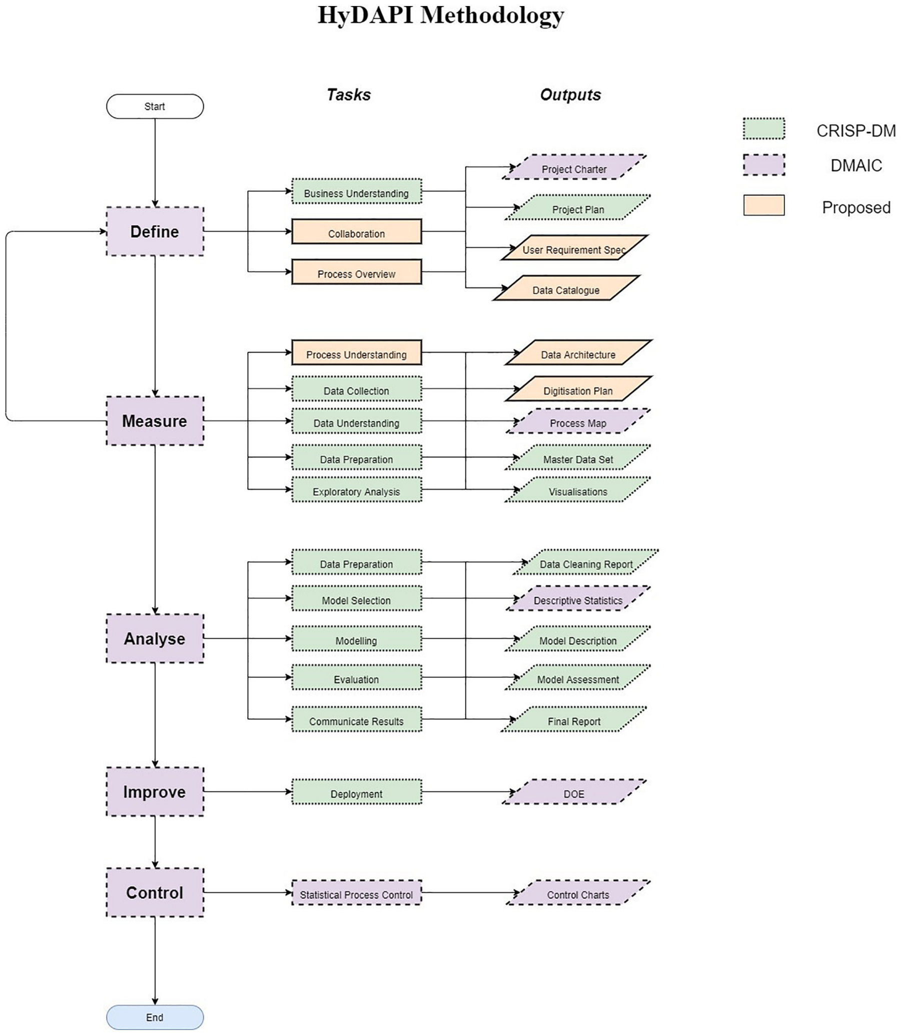
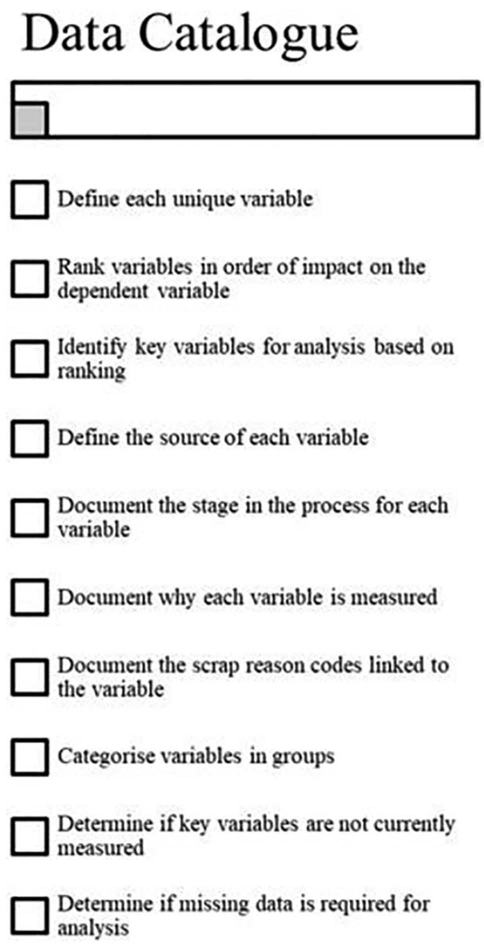


**Figure 4.**
HyDAPI methodology

tasks (discussed in detail later) contribute to ensuring there is collaboration between the manufacturing and data science disciplines. A **process overview** is the third task in the define phase of this methodology. It is similar to a product family analysis that is conducted as part of a value stream map (Vinet and Zhedanov, 2011). The author proposed the **process overview** task as it results in everyone involved in the project understanding the products manufactured in the process, the products that share common routing through the process, etc. The first output, the *project charter*, should contain detailed information, including specific key objectives that are attainable and realistic (Vinet and Zhedanov, 2011). Along with the objectives, the *project charter* should outline the current state, future state, benefits, key stakeholders, success measures, resources requirements, project lead and potential risks. Outlining the key stakeholders, resources required and the project lead in the *project charter* ensures that both the manufacturing and data science disciplines collaborate throughout the course of the project. The *project plan* should contain detailed actions, responsible resources, support resources and the timeline for the project. By clearly assigning each action with an owner, i.e. the responsible resource(s), this ensures that the manufacturing and data science disciplines are knitted together not only in the overall project but also within the smaller tasks of the project, helping to realise the **collaboration** task. The *user requirement specification* defines what the user wants to achieve, the intended audience, project scope and the assumptions and dependencies. The *user requirement specification* is a technical process originating from software engineering (Yue *et al.*, 2019). The *user requirement specification* output was proposed by the author to ensure that the end user, i.e. those looking to improve the manufacturing process agree with the data scientists what they would like to achieve from the project along with what is in or out of scope for the project. Yue *et al.* (2019) gives guidelines of how to define a user requirement specification based on ISA-95 standard. The *data catalogue*, which was proposed by the author, is the final output of the define phase. The steps to building a data catalogue are given in Figure 5. It uses an automation pyramid to define the data sources. It is a document containing all variables and associated metadata relating to the relevant manufacturing process. It is a useful document as it describes the scrap reasons that each variable is related to, if known, it describes the stage in the manufacturing process that it is related to and it defines key variables in the manufacturing process that are not currently recorded. To summarise, the tasks and outputs which have been proposed by the author for the define phase of this methodology are **collaboration**, **process overview**, *user requirement specification* and *data catalogue*. The purpose of the collaboration task is to ensure that both disciplines are working together throughout the course of the project. The second element proposed by the author is the **process overview task;** this task ensures the key process knowledge is shared with those performing the analysis. This is essential to guarantee that the analysis accurately represents the process. The next element proposed by the author is the *user requirement specification.* The purpose of the *user requirement specification* is to create an unambiguous document that outlines the desired outcomes of the project and constraints (Yue *et al.*, 2019). It prevents the end user having unrealistic expectations for the analysis; the data analysts are given a pragmatic path to follow and are informed of the composition and condition of the manufacturing data. The final element in the define phase that was proposed by the author was the *data catalogue* output. By creating a *data catalogue,* the data analysts can comprehend the array of variables for the manufacturing process, what they represent, the impact they have on the process and how to analyse them properly using the knowledge of the process.

### *5.2 Measure phase*

The next phase in the methodology is the measure phase; the purpose of this phase is to collect information relating to the process. The first task, **process understanding**, was proposed by the author because it is vital for the success of the project that all parties involved understand the data and the process being analysed. If actions are taken from inaccurate insights, both time and money are wasted, and trust is lost in the data and the data scientist. The **data collection** task is, as its name suggests, the process of gathering the initial dataset to be used for analysis. The **data understanding** task is conducted after the initial data collection and involves activities to identify data quality problems, to discover some minor insights and to understand the different subsets of the data (Chapman *et al.*, 2000). The **data preparation** task includes all steps taken to build the master data set from the initial raw data (Chapman *et al.*, 2000). It should be noted that the **data preparation** task can often take a significant amount of time. According to an article published in Forbes, data scientists spend 80% of their time preparing and managing data. The **data preparation**



Data Catalogue

□ Define each unique variable

□ Rank variables in order of impact on the dependent variable

□ Identify key variables for analysis based on ranking

□ Define the source of each variable

□ Document the stage in the process for each variable

□ Document why each variable is measured

□ Document the scrap reason codes linked to the variable

□ Categorise variables in groups

□ Determine if key variables are not currently measured

□ Determine if missing data is required for analysis

**Figure 5.**
Steps to creating a data catalogue

completed in the industrial demonstration was troublesome and inefficient. Therefore, the author proposes that for more efficient data preparation, a platform such as Jupyter notebook with a programming language like Python is used. Jupyter Notebooks are an excellent environment for cleaning data, visualising data, performing statistics with the data and creating machine-learning models. Using a programming language makes working with large datasets faster and it makes fixing errors or making changes much easier. If you alter your dataset, you can simply edit a few lines of code and re-run the analysis in a short time. The next task in the measure phase is **exploratory analysis.** This task uses visual tools to provide the analyst with a view of key metrics and measures within the organisation. Just like CRISP-DM, this methodology is an iterative process. It may occur to the project team when completing the **exploratory analysis**, that the data representing key input variables for the process is missing variables or not granular enough. If this is the case, the project team may circle back to the define phase of the project as seen in Figure 4. The first output in the measure phase is a *data architecture* diagram. The author proposed this output because data in manufacturing organisations are streamed from various different sources, both structured and unstructured (Belhadi *et al.*, 2019), and a data *architecture diagram* is a very helpful way to map the flow of data in the manufacturing system. A *data architecture* diagram created for the industrial use case is illustrated in Figure 5. It highlighted where certain data cannot be retrieved and to potential information sources that could be digitised. The next output in this phase is a *digitisation plan*. The author proposed the *digitisation plan* as it is needed to transform analogue data relating to the process that may currently be paper based documentation into digital format. Digitisation enables real-time collection of data which is crucial for effective processes (Kolberg and Zühlke, 2015). The next output in the measure phase is a *process map*. A *process map* allows you to see the data that is captured along the process, the different paths through the process for each product family and the process specifications that must be met. The next output of the measure phase is a *master data set*. The *master data set* is used for visualising, modelling and analysing the data. This dataset may need to be further modified before modelling, so there is another **data preparation** task in the analyse phase. The final output in the measure phase is data *visualisations*. This step acts as a guide in selecting which product family should be analysed in the next phase. Many large organisations produce a significant number of products; therefore, some initial work is required to identify the product family to be analysed. To conclude, the tasks and outputs which have been proposed by the author for the define phase of this methodology are **process understanding**, *Data Architecture* and *Digitisation Plan*. The author proposed the **process understanding** task because the industrial case study highlighted the importance of understanding the technical process when analysing the data. It is essential that the analysts understand the process in depth, to prevent misleading results. A *data architecture diagram* was proposed by the author to outline how information is collected for the process (see Figure 6). This diagram highlights potential areas along the process that need to be digitised. The digitisation plan was proposed by the author to transform the paper-based, analogue information collected along the process to digital format. The possible benefits associated with the digitisation of production has been getting attention from policy makers, for example, the European Commission has been focussing on ways to gather data related to the digitisation of production processes (Castelo-Branco *et al.*, 2019; European Commission, 2015). It is only possible to make improvements to processes (which is the goal of Six Sigma) if digital data are available (Seetharaman *et al.*, 2019). To achieve Industry 4.0 within the production of products, all process steps along the value stream must be digitised and interconnected (Leyh *et al.*, 2016), so that real-time information from the value stream ensures it is operating at its optimal capacity and efficiency (Nick *et al.*, 2019). Therefore, it is important to determine if data relating to the process is not easily retrievable and needs to be digitised. For it is the data, not the technology that will drive the transformation of the business (Klingenberg *et al.*, 2019; Vanauer *et al.*, 2015; Mohr and Hürtgen, 2018).

### *5.3 Analyse phase*

The analyse phase analyses the data that was gathered and digitised in the measure phase of this methodology with the purpose of identifying cause and effect relationships and causes for variation in the process. The first task, **data preparation**, involves cleaning the data to the level required for analysis, integrating data from different sources and formatting the data so that it is acceptable by the modelling tool (Chapman *et al.*, 2000). The **model selection** task refers to selecting the actual modelling technique to be used after reviewing the data and ensuring the assumptions required for the model are met. The **modelling** task involves splitting the data set into a training and testing set, fitting the model with the training set and completing predictions on the test set using the trained model. Models are built to explain and predict the data we observe. Model building is an iterative process with

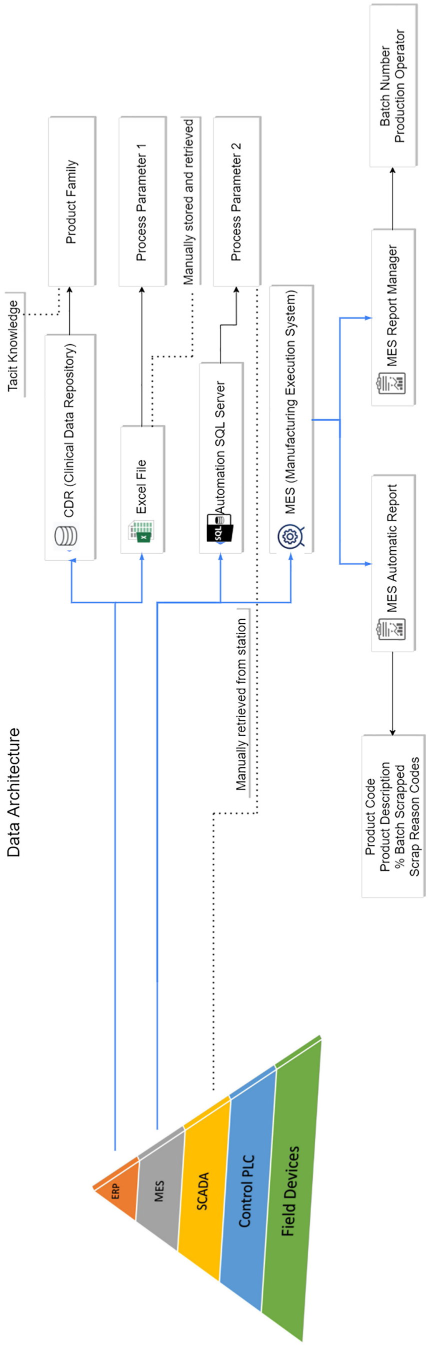


**Figure 6.**
Data architecture diagram

the aim being to get information from the data and align it with business requirements, constraints and feasibility (Finlay and Finlay, 2014). The **evaluation** task consists of evaluating the model using metrics such as a confusion matrix. A confusion matrix contains information about actual and predicted classifications including the number of correct negative predictions, incorrect positive predictions, incorrect negative predictions and correct positive predictions (Visa *et al.*, 2011). The final task in the analyse phase is to **communicate results** effectively to the senior management in the organisation. This is an important step in the process as presenting information in a way that people can effectively consume can be a challenge, and it is important so that decision makers are able to properly understand the data analysis and make concrete actions (Manyika *et al.*, 2011). The first output, the *data cleaning report* describes the steps taken to address the data quality (Chapman *et al.*, 2000). The *descriptive statistics* output is a key contribution of statistics and not taking them into account in data exploration and modelling restricts us to report values and parameter estimates without their corresponding variability (Weihs and Ickstadt, 2018). The *model description* is a document containing the model technique, model assumptions and the parameters in the model. The *model assessment* summarises the interpretation of the model(s) and ranks their quality in relation to each other (Chapman *et al.*, 2000). The *final report* is a document that includes the objectives and deliverables of the project and summarises the results of the analysis (Chapman *et al.*, 2000). All the tasks in the analyse phase originated from CRISP-DM, this highlights the need to incorporate data mining methods into traditional process improvement methodologies like Six Sigma because as businesses are transitioning to digital manufacturing, there is a need for new methods to analyse the growing complexity of manufacturing process data.

### *5.4 Improve phase*

The improve phase utilises the insights derived from the analyse phase and deploys them via actions to bring improvements to the process. The task in the improve phase is **deployment**. This involves taking the evaluation results from the analyse phase and concluding a strategy for **deployment** of the data mining results into the business (Chapman *et al.*, 2000). The output in this phase is a *Doe*. *Doe* stands for Design of Experiment and it originates from the Six Sigma DMAIC process. A *Doe* is used to statistically investigate the variables that influence a process and the resulting quality of products and services in an experimental setting (Gygi *et al.*, 2005). With a *Doe,* it is possible to understand the effect of changing multiple variables so that interaction relationships between variables can be seen.

### *5.5 Control phase*

The control phase is the final stage of the methodology, and it focuses on maintaining control of the process to ensure high quality products are consistently produced and quality costs are kept to a minimum. A recent study on the understanding of big data analytics for manufacturing processes states that only manufacturers that are able to analyse their manufacturing processes (with the exceedingly complex and excessive volume of manufacturing data) and proactively control their processes will survive in the next stage of advanced manufacturing (Belhadi *et al.*, 2019). To control the variation in the process, real-time data monitoring of the variables in the process is critical. The task in this phase is **statistical process control**. **Statistical process control** is part of the DMAIC process and it uses statistical techniques to monitor and control variation in processes (Gygi *et al.*, 2005). It can be used to stabilise out of control processes or to monitor the consistency of products and services. The output in this phase is *control charts*. *Control charts* originate from the DMAIC process, and they are the primary tool used as part of **statistical process**

**control**. They are a graphical tracking of a process input or output over time in relation to specified limits (Gygi *et al.*, 2005).

## 6. Conclusion

For companies to survive the transition to digital manufacturing and then to Industry 4.0, it is vital that they are capable of analysing their processes, forecast its optimal state and proactively control their processes (Krumeich *et al.*, 2014). The aim of data mining and analytics is to extract useful information from process data and use it to support decision-making leading to improved process performance (Krumeich *et al.*, 2014; Isaksson *et al.*, 2018). It is unanimously known that data-driven decision-making brought about by data mining and analytics leads to significant improvements in operational performance (Belhadi *et al.*, 2019). This study explored quality improvement and data mining methodologies that could be used for the digitisation of manufacturing processes to improve supply chain management performance using data-driven quality management. Upon assessment of methodologies in the process improvement and data science areas, Six Sigma and CRISP-DM were found to be the most suitable methodologies. The Six Sigma DMAIC and CRISP-DM methodologies were used in an industrial use case. However, the use of these two methodologies from the quality management and data science disciplines found that several gaps existed preventing the goal of using data-driven decision-making to improve the efficiency of the process and reduce the level of defective products produced. The author developed the HyDAPI methodology that combined the DMAIC and CRISP-DM methodologies along with further additions to fill these gaps. This proposed data-driven process improvement methodology acts as a guide for the digitisation of manufacturing processes to support decision-making leading to improved operation performance. In conclusion, the objective of the study was achieved in that the HyDAPI methodology was proposed that managers can utilise to digitise their operations, enabling data-driven quality management of their manufacturing operations and achieve improved supply chain management performance. This also answers the research question "How can industrial practitioners use existing quality management concepts in the digitisation of their supply chain operations to achieve data-driven quality management and improved supply chain performance?".

### *6.1 Practical implications*

The literature has stated that it is not clear how industry practitioners can implement digitisation of manufacturing processes in the era of Industry 4.0 and use data mining techniques to achieve data-driven quality management and improved supply chain performance (Petersen *et al.*, 2017; Johansson *et al.*, 2019b; Akter *et al.*, 2016). This research work reviewed the existing quality improvement and data mining methodologies for their potential use in achieving data-driven quality management through the digitisation of supply chain operations. A review of the literature was conducted to determine the most suitable methodologies from both the quality management and data science disciplines. Valuable practical learnings borne out of the implementation of these methodologies were used to develop the HyDAPI methodology. This methodology offers a practical step by step approach for practitioners in industry to follow in transforming their traditional processes to digital to achieve data-driven quality management. The methodology could provide insights for enabling practitioners to understand the necessity to combine data science as well as process expertise when digitally transitioning their supply chain operations. New roles may be developed in industry to facilitate the merging of these two disciplines to digitise the operations and achieve data-driven quality improvement. The methodology could aid managers to understanding and conceptualise the integration of their existing quality management concepts with data mining processes in adapting to Industry 4.0.

### 6.2 Limitations and future research directions

The review of existing methodologies in the quality management and data science is not comprehensive, and there are possibly other methodologies from the data science and quality management disciplines which have not been included in the literature review of this study. There are also mixed reviews of the use of the chosen methodologies Six Sigma DMAIC and CRISP-DM. For instance, Six Sigma is not considered suitable for highly innovative, fast-paced processes (Basios and Loucopoulos, 2017). However, using the defined criteria to compare the methodologies, these are deemed the most suitable methodologies for the digitisation of operations in the supply chain to achieve data-driven quality management. Proposals for future research are to implement of the proposed HyDAPI methodology in an in-depth industrial case study to demonstrate and test its benefits and limitations. Findings from the full implementations of the HyDAPI methodology could enable further revision and enhancement of the methodology, as well as validate its benefits in practical implementation.

## References

Akter, S., Fossa Wamba, S., Gunasekaran, A., Dubey, R. and Childe, S.J. (2016), "How to improve firm performance using big data analytics capability and business strategy alignment?", *International Journal of Production Economics*, Vol. 182, pp. 113-131, doi: 10.1016/j.ijpe.2016.08.018.

Andersson, R., Eriksson, H. and Torstensson, H. (2006), "Similarities and differences between TQM, six sigma and lean", *TQM Magazine*, Vol. 18 No. 3, pp. 282-296, doi: 10.1108/09544780610660004.

Azevedo, A. and Santos, M.F. (2008), "KDD, SEMMA and CRISP-DM: a parallel overview", *IADIS European Conference Data Mining*, pp. 182-185.

Bag, S., Yadav, G., Wood, L.C., Dhamija, P. and Joshi, S. (2020a), "Industry 4.0 and the circular economy: resource melioration in logistics", *Resources Policy*, Vol. 68 May, 101776, doi: 10.1016/j.resourpol.2020.101776.

Bag, S., Wood, L.C., Xu, L., Dhamija, P. and Kayikci, Y. (2020b), "Big data analytics as an operational excellence approach to enhance sustainable supply chain performance", *Resources, Conservation and Recycling*, Vol. 153 November 2019, 104559, doi: 10.1016/j.resconrec.2019.104559.

Bag, S., Wood, L.C., Mangla, S.K. and Luthra, S. (2020c), "Procurement 4.0 and its implications on business process performance in a circular economy", *Resources, Conservation and Recycling*, Vol. 152 September 2019, 104502, doi: 10.1016/j.resconrec.2019.104502.

Bang, S.H., Ak, R., Narayanan, A., Lee, Y.T. and Cho, H. (2019), "A survey on knowledge transfer for manufacturing data analytics", *Computers in Industry*, Vol. 104 March 2018, pp. 116-130, doi: 10.1016/j.compind.2018.07.001.

Basios, A. and Loucopoulos, P. (2017), "Six sigma DMAIC enhanced with capability modelling", *Proceedings-2017 IEEE 19th Conference on Business Informatics, CBI 2017*, Vol. 2, pp. 55-62, doi: 10.1109/CBI.2017.70.

Belhadi, A., Zkik, K., Cherrafi, A., Yusof, S.M. and El Fezazi, S. (2019), "Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies", *Computers and Industrial Engineering*, Vol. 137 No. 106099, doi: 10.1016/j.cie.2019.106099.

Buer, S.V., Strandhagen, J.O. and Chan, F.T.S. (2018), "The link between industry 4.0 and lean manufacturing: mapping current research and establishing a research agenda", *International Journal of Production Research*, Vol. 56 No. 8, pp. 2924-2940, doi: 10.1080/00207543.2018.1442945.

Castelo-Branco, I., Cruz-Jesus, F. and Oliveira, T. (2019), "Assessing industry 4.0 readiness in manufacturing: evidence for the European union", *Computers in Industry*, Vol. 107, pp. 22-32, doi: 10.1016/j.compind.2019.01.007.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000), *Step-by-step Data Mining Guide*, SPSS, Chicago, IL.

Chong, D. and Shi, H. (2015), "Big data analytics: a literature review", *Journal of Management Analytics*, Vol. 2 No. 3, pp. 175-201, doi: 10.1080/23270012.2015.1082449.

Clancy, R., Ahern, M., Sullivan, D.O. and Bruton, K. (2020), "Systematic mapping study of digitization and analysis of manufacturing data", *International Journal of Industrial and Systems Engineering*, Vol. 14 No. 9, pp. 717-731.

Court, D. (2015), *Marketing and Sales Big Data, Analytics and the Future of Marketing and Sales*, McKinsey and Company, Chicago, pp. 8-14, doi: 10.1016/j.cie.2019.106099.

Dahlgaard, J.J., Khanji, G.K. and Kristensen, K. (1997), *Fundamentals of Total Quality Management*, 1st ed., Routledge, London.

De Mast, J. and Lokkerbol, J. (2012), "An analysis of the six sigma DMAIC method from the perspective of problem solving", *International Journal of Production Economics*, Vol. 139 No. 2, pp. 604-614, doi: 10.1016/j.ijpe.2012.05.035.

Dhamija, P. and Bag, S. (2020), "Role of artificial intelligence in operations environment: a review and bibliometric analysis", *TQM Journal*, Vol. 32 No. 4, pp. 869-896, doi: 10.1108/TQM-10-2019-0243.

Elg, M., Birch-Jensen, A., Gremyr, I., Martin, J. and Melin, U. (2020), "Digitalisation and quality management: problems and prospects", *Production Planning and Control*, Vol. 31, pp. 1-14, doi: 10.1080/09537287.2020.1780509.

Eskildson, L. (1994), "'Improving the odds of TQM's success'", *Quality Progress*, Vol. 27 No. 4, doi: 10.1108/02656710310500815.

European Commission (2015), "Monitoring the digital economy and society 2016–2021", *European Commission DG Communications Networks, Content and Technology*, p. 52.

Fayyad, U., Piatetsky-Shapiroo, G. and Smyth, P. (1996), "From data mining to knowledge discovery in databases", *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 17 No. 3, pp. 637-648, doi: 10.1007/978-3-319-18032-8_50.

Finlay, S. and Finlay, S. (2014), "How to build a predictive model", *Predictive Analytics, Data Mining and Big Data*, 1st ed., Palgrave Macmillan, pp. 157-178, ISBN 978-1-137-37928-3, doi: 10.1057/9781137379283_8.

Foley, K. (2004), *Five Essays on Quality Management*, SAI Global, Sydney.

Ge, Z., Song, Z., Ding, S.X. and Huang, B. (2017), "Data mining and analytics in the process industry: the role of machine learning", *IEEE Access*, Vol. 5, pp. 20590-20616, doi: 10.1109/ACCESS.2017.2756872.

Gygi, C., DeCarlo, N. and Williams, B. (2005), in Drenth, T.S. (Ed.), *Six Sigma for Dummies*, Wiley Publishing.

Huber, W. (2016), "Digital business processes", *Industry 4.0 in Automobile Production: A Practice Book*, 1st ed., Springer Viewer, pp. 245-258, doi: 10.1007/978-3-658-12732-9.

Isaksson, A.J., Harjunkoski, I. and Sand, G. (2018), "The impact of digitalization on the future of control and operations", *Computers and Chemical Engineering*, Vol. 114, pp. 122-129, doi: 10.1016/j.compchemeng.2017.10.037.

Johansson, P.E.C., Malmsköld, L., Fast-Berglund, Å. and Moestam, L. (2019), "Challenges of handling assembly information in global manufacturing companies", *Journal of Manufacturing Technology Management*, Vol. 31 No. 5, doi: 10.1108/JMTM-05-2018-0137.

Klefsjö, B., Wiklund, H. and Edgeman, R.L. (2001), "Six sigma seen as a methodology for total quality management", *Measuring Business Excellence*, Vol. 5 No. 1, pp. 31-35, doi: 10.1108/13683040110385809.

Klingenberg, C.O., Borges, M.A.V. and Antunes, J.A.V. (2019), "Industry 4.0 as a data-driven paradigm: a systematic literature review on technologies", *Journal of Manufacturing Technology Management*, Vol. 32 No. 3, doi: 10.1108/JMTM-09-2018-0325.

Knowles, G. (2011), *Six Sigma*, Ventus Publishing ApS, London.

Kolberg, D. and Zühlke, D. (2015), "Lean automation enabled by industry 4.0 technologies", *IFAC-PapersOnLine*, Vol. 28 No. 3, pp. 1870-1875, doi: 10.1016/j.ifacol.2015.06.359.

Kristoffersen, E., Aremu, O.O., Blomsma, F., Mikalef, P. and Li, J. (2019), "Exploring the relationship between data science and circular economy: an enhanced CRISP-DM process model", *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11701 LNCS June, pp. 177-189, doi: 10.1007/978-3-030-29374-1_15.

Krumeich, J., Jacobi, S., Werth, D. and Loos, P. (2014), "Big data analytics for predictive manufacturing control - a case study from process industry", *Proceedings-2014 IEEE International Congress on Big Data, BigData Congress 2014*, pp. 530-537, doi: 10.1109/BigData.Congress.2014.83.

Leyh, C., Bley, K., Schaffer, T. and Forstenhausler, S. (2016), "SIMMI 4.0-a maturity model for classifying the enterprise-wide it and software landscape focusing on Industry 4.0", *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016*, Institute of Electrical and Electronics Engineers, pp. 1297-1302, doi: 10.15439/2016F478.

Manyika, J., Chui, M., Brown, B., Jacques, B., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011), *Big Data: The Next Frontier for Innovation, Competition and Productivity*, McKinsey Global Institute, New York.

Mayr, A., Weigelt, M., Kühl, A., Grimm, S., Erll, A., Potzel, M. and Franke, J. (2018), "Lean 4.0-A conceptual conjunction of lean management and Industry 4.0", *Procedia CIRP*, Vol. 72, pp. 622-628, doi: 10.1016/j.procir.2018.03.292.

McCue, C. (2015), "Process models for data mining and predictive analysis", *Data Mining and Predictive Analysis*, 2nd ed., pp. 51-74, doi: 10.1016/b978-0-12-800229-2.00004-3.

Middleton, N.T. (1996), "Applying the continuous improvement process to engineering design in the laboratory", *Proceedings - Frontiers in Education Conference*, Vol. 3, pp. 1429-1433, doi: 10.1109/fie.1996.568533.

Mohr, N. and Hürtgen, H. (2018), *Achieving Business Impact with Data A Comprehensive Perspective on the Insights Value Chain*, McKinsey Quarterly, New York.

Mrugalska, B. and Wyrwicka, M.K. (2017), "Towards lean production in industry 4.0", *Procedia Engineering*, Vol. 182, pp. 466-473, doi: 10.1016/j.proeng.2017.03.135.

Nick, G., Szaller, Á., Bergmann, J. and Várgedő, T. (2019), "Industry 4.0 readiness in Hungary: model, and the first results in connection to data application", *IFAC-PapersOnLine*, Vol. 52 No. 13, pp. 289-294, doi: 10.1016/j.ifacol.2019.11.185.

Palacios, H.J.G., Toledo, R.A.J., Pantoja, G.A.H. and Navarro, Á.A.M. (2017), "A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change", *Advances in Science, Technology and Engineering Systems*, Vol. 2 No. 3, pp. 598-604, doi: 10.25046/aj020376.

Petersen, N., Halilaj, L., Grangel-González, I., Lohmann, S., Lange, C. and Auer, S. (2017), "Realizing an RDF-based information model for a manufacturing company – a case study", in, d'Amato, C. *et al.* (Eds), *International Semantic Web Conference*, Springer International Publishing (Lecture Notes in Computer Science), Cham. doi: 10.1007/978-3-319-68204-4.

Pyvovar, N., Yamkovyi, K. and Vechirko, M. (2019), "Data science project management methdologies", Medium Corporation, San Francisco, available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb.

Rüttimann, B.G. and Stöckli, M.T. (2016), "'Lean and industry 4.0—twins, partners, or contenders? A due clarification regarding the supposed clash of two production systems'", *Journal of Service Science and Management*, Vol. 09 No. 06, pp. 485-500, doi: 10.4236/jssm.2016.96051.

Schafer, F., Zeiselmair, C., Becker, J. and Otten, H. (2018), "Synthesizing CRISP-DM and quality management: a data mining approach for production processes", *2018 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD, 2018*, pp. 190-195, doi: 10.1109/ITMC.2018.8691266.

Schafer, F., Schwulera, E., Otten, H. and Franke, J. (2019), "From descriptive to predictive six sigma: machine learning for predictive maintenance", *Proceedings - 2019 2nd International Conference on Artificial Intelligence for Industries, AI4I, 2019*, pp. 35-38, doi: 10.1109/AI4I46381.2019.00017.

Schroeder, R.G., Linderman, K., Liedtke, C. and Choo, A.S. (2008), "Six sigma: definition and underlying theory", *Journal of Operations Management*, Vol. 26 No. 4, pp. 536-554, doi: 10.1016/j.jom.2007.06.007.

Seetharaman, A., Patwa, N., Saravanan, A.S. and Sharma, A. (2019), "Customer expectation from industrial internet of things (IIOT)", *Journal of Manufacturing Technology Management*, Vol. 30 No. 8, pp. 1161-1178, doi: 10.1108/JMTM-08-2018-0278.

Shafique, U. and Qaiser, H. (2014), "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)", *International Journal of Innovation and Scientific Research*, Vol. 12 No. 1, pp. 217-222.

Sharma, A. and Pandey, H. (2019), "Big data and analytics in industry 4.0", *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*, pp. 57-72.

Swarnakar, V. and Vinodh, S. (2016), "Deploying Lean Six Sigma framework in an automotive component manufacturing organization", *International Journal of Lean Six Sigma*, Vol. 7 No. 3, pp. 267-293, doi: 10.1108/IJLSS-06-2015-0023.

Telukdarie, A., Buhulaiga, E., Bag, S., Gupta, S. and Luo, Z. (2018), "Industry 4.0 implementation for multinationals", *Process Safety and Environmental Protection*, Vol. 118, pp. 316-329, doi: 10.1016/j.psep.2018.06.030.

Vanauer, M., Bohle, C. and Hellingrath, B. (2015), "Guiding the introduction of big data in organizations: a methodology with business- and data-driven ideation and enterprise architecture management-based implementation", *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2015-March, pp. 908-917, doi: 10.1109/HICSS.2015.113.

Vinet, L. and Zhedanov, A. (2011), "A 'missing' family of classical orthogonal polynomials", *Journal of Physics A: Mathematical and Theoretical*, Vol. 44 No. 8, doi: 10.1088/1751-8113/44/8/085201.

Visa, S., Ramsay, B., Ralescu, A. and van der Knaap, E. (2011), "Confusion matrix-based feature selection", *Proceedings of the Twenty second Midwest Artificial Intelligence and Cognitive Science Conference*, pp. 120-127.

Wagner, T., Herrmann, C. and Thiede, S. (2017), "Industry 4.0 impacts on lean production systems", *Procedia CIRP*, Vol. 63, pp. 125-131, doi: 10.1016/j.procir.2017.02.041.

Weihs, C. and Ickstadt, K. (2018), "Data Science: the impact of statistics", *International Journal of Data Science and Analytics*, Vol. 6 No. 3, pp. 189-194, doi: 10.1007/s41060-018-0102-5.

Weill, P. and Woerner, S.L. (2015), *Thriving in an Increasingly Digital Ecosystem*, MIT SLOAN Management Review, Massachusetts.

Yue, L., Niu, P. and Wang, Y. (2019), "Guidelines for defining user requirement specifications (URS) of manufacturing execution system (MES) based on ISA-95 standard", *Journal of Physics: Conference Series*, Vol. 1168 No. 3, doi: 10.1088/1742-6596/1168/3/032065.

**Corresponding author**
Rose Clancy can be contacted at: rosieclancy97@gmail.com