UTS UNIVERSITY OF TECHNOLOGY SYDNEY

A Process Mining Approach for Production Flow Analysis

by Laura Tomidei

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Dr.-Ing. Jochen Deuse, Dr. Nathalie Sick, Dr.-Ing. Matthias Guertler, Dr. Luke Mathieson

University of Technology Sydney
Faculty of Engineering and IT, School of Mechanical and Mechatronic Engineering

August 2024

Certificate of Original Authorship

I, Laura Tomidei, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Mechanical and Mechatronic Engineering (Faculty of Engineering and IT) at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note: Signature removed prior to publication.

Date: 20/08/2024

Acknowledgements

Throughout my PhD journey, I have been fortunate to receive the support of many individuals, and I would like to take this opportunity to express my gratitude to them.

I would like to thank my principal supervisor, Jochen Deuse, for suggesting the research topic, providing me with invaluable research data, and giving me academic freedom.

I would like to express my deepest gratitude to my co-supervisors, Nathalie Sick and Matthias Guertler, who encouraged me to pursue a PhD and have been my mentors since well before the start of this journey. Their guidance and continuous support, especially during challenging times, have been a source of strength throughout this journey. Their invaluable mentorship and their encouragement to pursue academic opportunities within the PhD and in related areas beyond it, have shaped me into what I believe a balanced academic. I owe them a tremendous thank you for their support throughout this academic journey, which has been instrumental in helping me achieve what I have so far.

I would also like to thank my co-supervisor, Luke Mathieson, who although joined the supervisory panel later in the candidature, provided me with invaluable support on some of the most difficult technical aspects of this work.

I would like to express my appreciation to all the colleagues and friends at the School of Mechanical and Mechatronic Engineering who have always made me feel part of the team and given me constant support.

Finally, I would like to thank my family who, although are physically distant, I feel constantly close to. The section below is dedicated to them.

Un ringraziamento speciale va alla mia famiglia in Italia. Mi sento molto fortunata ad avere una famiglia alle spalle che mi incoraggia costantemente a seguire i miei sogni, anche se questi sono dall'altra parte del mondo. La cosa che mi rende più felice nel completare questo percorso accademico e la consapevolezza di rendervi orgogliosi. Nonostante la lontananza, sento il vostro constante affetto e incoraggiamento. Senza il vostro supporto, niente di tutto ciò sarebbe stato possibile e ve ne sarò grata per sempre.

Laura

August 2024

Keywords

Industrial Data Science, Industry 4.0, Group Technology, Machine Learning, Process Mining, Production Flow Analysis, Production Planning and Control, Value Stream Analysis, Value Stream Identification.

Abstract

Lean Management principles and methods have been adopted by manufacturing companies for decades, as they enable the design, planning, and control of efficient production systems. One of the key principles of Lean Management and an essential requirement for the effective application of Production Planning and Control tasks is the identification of value streams in production. Generally defined as the steps required to bring a product to the customer, value streams are specific to a product or product family where variants are regarded as one representative product.

As a key enabling step for future improvements, identifying value streams allows the effective application of methods such as Value Stream Mapping, production levelling, and push production. In order to identify value streams, having a clear understanding of production flows is key. In practice, this task can be challenging, particularly for high-mix low volume companies, for which value streams may be composed of hundreds of products and parts.

To analyse the complexity of production flows and identify the similarity between products and resources, Burbidge first introduced the principles for Production Flow Analysis in 1970s. Intended to simply the material flow and re-organise factories into groups of machines that complete all parts they make, the method enables the identification of product families and related processing steps. As such, the principles of Production Flow Analysis can be used for value stream identification. However, as the complexity of production systems has grown significantly over the last decade, the direct applicability of these principles has reached its limits. At the same time, the increasing adoption of emerging technologies in industrial contexts has unlocked new opportunities. In the age of Industry 4.0, or fourth industrial revolution, advanced digitalisation is increasingly integrated into production systems, leading to greater availability of information. To transform this data into fact-based insights that can assist Production Planning and Control decisions, Data Analytics techniques are key. In particular, recent developments in Process Mining have enabled the discovery, analysis, and improvement of processes in the manufacturing domain and beyond.

By exploiting these advancements and building on the principle of Production Flow Analysis, this research provides a method for automatically identifying value streams from production data. Through the definition and analysis of process flows and their homogeneity, the method uses Process Mining and Machine Learning techniques to structure production processes into value streams and visualise the associated process models. The effectiveness of the method is validated using a case study. Since the characteristics of data collection and management systems used by companies may vary across the industry, this method also takes into consideration use cases with different levels of data quality.

By providing a solution for effectively using production data and accurately identifying value streams, the method represents a novel contribution to both academia and industry. From a theoretical perspective, this work builds on the fundamental principles of Production Flow Analysis and further advances them using Process Mining techniques to provide a conceptual method for automated value stream identification. From a practical perspective, this work addresses practical challenges by enabling visibility and transparency in production and facilitating the effective application of Production Planning and Control techniques in modern manufacturing environments.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

IE	Industrial Engineering
PPC	Production Planning and Control
PFA	Production Flow Analysis
GT	Group Technology
PM	Process Mining
DRM	Design Research Methodology
IoT	Internet of Things
CPS	Cyber Physical Systems
ML	Machine Learning
ANN	Artificial Neural Networks

List of Publications

Tomidei, L., Sick, N. & Mathieson, L. (2024). Data-Driven Value Stream Analysis Using Process Mining And Machine Learning. In *51st International Conference on Computers and Industrial Engineering (CIE51)*. Sydney, Australia.

Tomidei, L., Sick, N., Deuse, J. & Guertler, M. (2023). Extracting Key Value Streams using Process Mining and Machine Learning. In *IEEE Conference on Engineering Informatics*. *2023 IEEE Engineering Informatics*, 1–7. https://doi.org/10.1109/IEEECONF58110.2023.10520644

Tomidei, L., Sick, N., Deuse, J., & Clemon, L. (2022). Production Flow Analysis in the Era of Industry 4.0 : How Digital Technologies can Support Decision-Making in the Factory of the Future. In *2022 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 1-15). Piscataway, USA: IEEE. doi:10.23919/picmet53225.2022.9882711

Other Publications

Wambsganss, A., Tomidei, L., Sick, N., Salomo, S., & Miled, E. B. (2024). Machine learning-based method to cluster a converging technology system: The case of printed electronics. *World Patent Information*, *78*, 102301.

Tomidei, L., Guertler, M., Sick, N., Paul, G., Carmichael, M. (2024). Design Principles for Safe Human Robot Collaboration. INTERACTION DESIGN & ARCHITECTURE – IxD&A Journal, Special Issue.

Tomidei, L., Sick, N., Guertler, M., Schallow, J., Lenze, D., & Deuse, J. (2023). Dynamic value stream mapping: How Industry 4.0 can help us to learn to see better. In *9th Changeable, Agile, Reconfigurable and Virtual Production Conference*. Bologna, Italy.

Wambsganss, A., Tomidei, L., Sick, N., Bröring, S., Salomo, S., & Schultz, C. (2023). Machine-based anticipation of converging technology systems: The case of printed electronics. In *International Society for Professional Innovation Management*. Ljubljana, Slovenia.

Rokoss, A., Syberg, M., Tomidei, L., Huelsing, C., Deuse, J., & Schmidt, M. (2023). Case study on delivery time determination using a machine learning approach in small batch production companies. *JOURNAL OF INTELLIGENT MANUFACTURING*, 22 pages. doi:10.1007/s10845-023-02290-2

Guertler, M., Tomidei, L., Sick, N., Carmichael, M., Paul, G., Wambsganss, A., . . . Hussain, S. (2023). WHEN IS A ROBOT A COBOT? MOVING BEYOND MANUFACTURING AND ARM-BASED COBOT MANIPULATORS. *Proceedings of the Design Society*, *3*, 3889-3898. doi:10.1017/pds.2023.390

Guertler, M., Carmichael, M., Paul, G., Sick, N., Tomidei, L., Hernandez Moreno, V., . . . Hussain, S. (2022). *Guidelines for the Safe Collaborative Robot Design*

and Implementation. NSW Government: Centre for Work Health and Safety: NSW Government: Centre for Work Health and Safety. Retrieved from https://www.centreforwhs.nsw.gov.au/

Paul, G., Tomidei, L., Sick, N., Guertler, M., Carmichael, M., & Wambsganss, A. (2022). Guidelines for Safe Collaborative Robot Design and Implementation. In *Guidelines for Safe Collaborative Robot Design and Implementation*. Sydney.

Frijat, L., Tomidei, L., Guertler, M., & Sick, N. (2022). Collaborative Robotics: A new work health and safety risk assessment for a novel technology. In *Asia Pacific Occupational Safety & Health Organization*. Melbourne.

Tomidei, L., Sick, N., Guertler, M., Frijat, L., Carmichael, M., Paul, G., . . . Hussain, S. (2022). BEYOND TECHNOLOGY - THE COGNITIVE AND ORGANISATIONAL IMPACTS OF COBOTS. In *Australasian Conference on Robotics and Automation* Vol. 2022.

Clemon, L., Guertler, M., Tomidei, L., & Edwards, R. (2021). *Additive manufacturing opportunities for Australia's agriculture, fisheries and forestry sectors* (21-089). Wagga Wagga, NSW: AgriFutures - National Rural Issues. Retrieved from https://www.agrifutures.com.au/

Lammers, T., Tomidei, L., & Trianni, A. (2019). Towards a Novel Framework of Barriers and Drivers for Digital Transformation in Industrial Supply Chains. In *2019 Portland International Conference on Management of Engineering and Technology (PICMET)*. Portland: IEEE. doi:10.23919/picmet.2019.8893875

Lammers, T., Tomidei, L., & Regatierri, A. (2018). What Causes Companies to Transform Digitally? An Overview of Drivers for Australian Key Industries. In *Portland International Conference on Management of Engineering and Technology*. Honolulu, HI, USA: IEEE. doi:10.23919/PICMET.2018.8481810

Chapter 1: Introduction

In manufacturing, companies are constantly striving for continuous improvements in productivity, quality, and level of service (Ejsmont et al., 2020a). For decades, Lean Management has enabled enterprises to achieve these goals through methods and techniques that focus on value-added activities while reducing different forms of waste in production. One of the key principles of Lean Management is identifying and selecting value streams in production (Womack & Jones, 1997). Generally, value streams are defined as all the actions needed to bring a product (or a family of products) to the customer (Rother & Shook, 2003; Womack & Jones, 1997). Although identifying value streams may not always be simple, particularly for high mix-low volume companies, whose value streams can be composed of hundreds of parts and products (Braglia et al., 2006), it is an essential task for the effective application of Production Planning and Control techniques such as layout design (Burbidge, 1991), value stream analysis (Rother & Shook, 2003), scheduling (Bohnen et al., 2011), and line balancing (Deuse et al., 2013).

The task of value streams identification requires distinguishing homogenous groups of products and resources according to their similarity. In other words, value streams identification closely relates to Group Technology (GT), which is an engineering and manufacturing approach that identifies the similarity of products as well as the equipment or processing steps used to make them (Hameri, 2011). Group Technology is essential for optimising production processes, as organising processes and structures based on homogenous objects and resources is often more effective and efficient (Deuse et al., 2022). Originally, the goal of Group Technology was to transfer the benefits of economies of scale to batch and job shop production (Deuse et al., 2013). As such, it represents a method of factory organisation where organisational units or groups complete all the products they make (Burbidge, 1991). To achieve this, several approaches have been introduced over the years. While early methods relied on classification and coding systems that group processes and machines based on products' similarity in shape (Mitrofanov, 1961), in 1970s Burbidge introduced Production Flow Analysis, a technique that uses component

process routes to find groups of processing steps and associated product families (Burbidge, 1991).

In the last few decades, as processes are becoming more complex, routing flexibility and variability have also increased, ultimately making the task of value stream identification more challenging. Mass customisation trends have impacted the production characteristics of many enterprises, with higher number of variants, and lower volumes of identical parts (Deuse et al., 2013). At the same time, the paradigm shift in industrial applications driven both by these application-pull and technology-push factors described as “fourth industrial revolution”, or “Industry 4.0”, has unlocked new opportunities. Thus, advanced digitalisation combined with Internet technologies and “smart” objects are being increasingly integrated into the production systems. “Smart factories” are characterised by production systems that are equipped with sensors, actors, and autonomous systems (Lasi et al., 2014). The integration of the physical system with the software system leads to an increasing amount of information that can assist Production Planning and Control decisions and ultimately improve visibility, transparency, predictive capability, and adaptability in production processes (Schuh et al., 2020).

As a result, existing research has investigated the synergies between many traditional approaches and Industry 4.0 technologies. However, the potential of supporting Production Flow Analysis with new data-driven techniques has received limited attention. In particular, among promising modern techniques, Process Mining (PM) is considered a useful tool for addressing Data Analytics tasks with challenges arising from process complexity, due to its ability to discover process models from event data and provide fact-based insights (Halaska & Sperka, 2018; Knoll et al., 2019). As such, Process Mining is considered “the logical next step in the development of Group Technology” (Deuse et al., 2022, p. 6).

Therefore, the goal of this thesis is the following:

Developing a Process Mining-based approach to identify key value streams and enable the effective application of Production Planning and Control techniques in complex environments.

To address this research question, this work follows the key steps outlined in Design Research Methodology (DRM) proposed by Blessing & Chakrabarti (2009), an established framework in engineering design. In line with the purpose of the DRM, this work aims to develop a support system whose goal is to solve a practically relevant research problem while simultaneously providing a theoretical contribution.

While the main steps of the DRM listed below may appear linear, in practice this methodology involves iterations that increase the understanding of the problem as well as parallel executions (see Figure 1-1). The structure of this thesis follows the key steps of the DRM. While Figure 1-1 provides and overview of the overall methodology and structure, detailed methodology sections are included in each chapter.

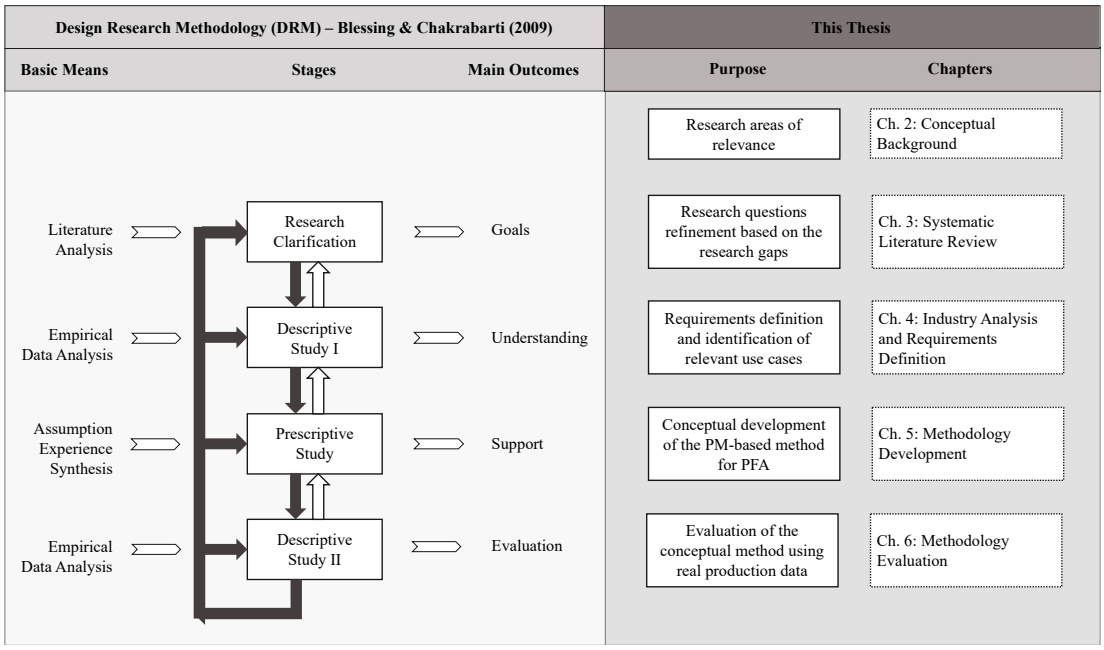


Figure 1-1: Research Design Method based on Blessing & Chakrabarti (2009)

In the background section, an overview of the areas of relevance and contribution is provided by illustrating how this research integrates key elements from Group Technology, Industry 4.0, and Lean Production. Driving process analysis and improvements for decades, Lean Production methods require identifying value streams as a first step, thus highlighting the importance of the task. Group

Technology enabled by Production Flow Analysis represents how this task can be achieved, while Industry 4.0 capabilities and technologies provide a solution for handling complexity and increasing amounts of data in complex environments. Among these research areas, two are considered essential for the development of this work, namely Process Mining and Production Flow Analysis.

Based on this consideration, the systematic literature review provides a thorough investigation of the essential areas Process Mining and Production Flow Analysis by evaluating potential synergies and highlighting research gaps. The goal is to present a detailed understanding of the research gaps and validate the research objective/question (Blessing & Chakrabarti, 2009).

The descriptive study I aims to define relevant use cases. To this end, a combination of empirical data analysis and literature analysis is used to understand data requirements and maturity levels. In fact, companies' level of digital maturity and the characteristics of their data collection and storage systems often vary across the industry. Thus, use cases are derived based on the possible data maturity levels and their corresponding data requirements.

The prescriptive study develops a method for identifying value streams automatically. The method builds on existing procedures and frameworks to develop a new solution for all use cases defined in the previous phase.

Finally, the descriptive study II evaluates the conceptual methodology by validating it with a case study to demonstrate the applicability to practical contexts. Using exemplary use cases is an established approach in the broader field of Machine Learning, as it enables insights with practical relevance (McCutcheon & Meredith, 1993).

Chapter 2: Conceptual Background

The following sections present the conceptual background on which this research has been developed. The Areas of Relevance and Contribution Diagram shown in Figure 2-1 clarifies the foundations on which this research is based as well as the areas of contribution (Blessing & Chakrabarti, 2009).

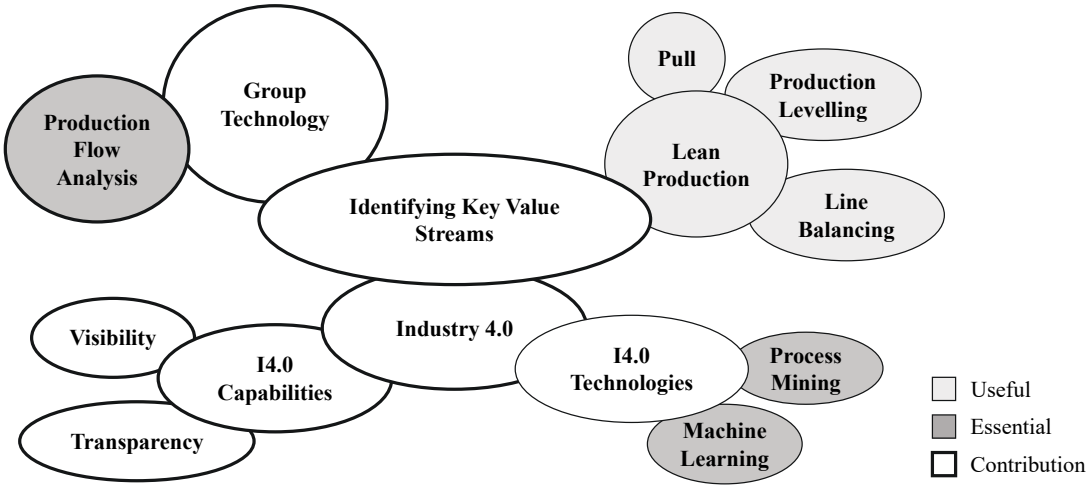


Figure 2-1: Areas of Relevance and Contribution diagram

There are four main research areas involved in the identification of key value streams, namely Lean Production, Group Technology, Industry 4.0 capabilities, and Industry 4.0 technologies. The following sections provide an overview of each area, with a focus on the relation to the task of identifying value streams. First, an overview of value streams identification highlights the importance of this task in many IE applications, particularly for Lean techniques (2.1). Second, Production Flow Analysis is presented as an enabling technique for planning Group Technology and identifying key value streams (2.2). Third, an overview of Industry 4.0 illustrates the potential opportunities and capabilities that technologies can unlock in industrial environments (2.3). Finally, an overview of Processes Mining is presented together with the opportunities it unlocks in relation to grasping process complexity (2.4).

2.1 THE IMPORTANCE OF VALUE STREAM IDENTIFICATION IN LEAN PRODUCTION

A value stream is defined as "all the actions (both value-creating and nonvalue-creating) required to bring a product through the main flows essential to every

product: (1) the production flow from raw material into the arms of the customer, and (2) the design flow from concept to launch" (Rother & Shook, 2003). The key requirement for value stream representation and analysis is focusing on one product family (Rother & Shook, 2003). Therefore, for the purpose of production analysis and design, a value stream is identified as the material flow of a single product family. Within a product family, all variants are treated as one representative product (Erlach, 2013). The identification of value streams is a key step in production optimisation. The underlaying idea is that by adopting the value stream perspective it is possible to transform the production into a value creating flow (Erlach, 2013). In the context of Lean Production, that started in 1950s with the Toyota Production System (TPS), following the book "The Machine that Changed the World", Womack & Jones (1997) defined the key principles of Lean Thinking, namely specify value, identify and map the value stream, make the value flow, apply pull, and pursue perfection. The goal of these principles is identifying and removing sources of waste in production, which could come in different forms. While *muda* refers to 8 types of waste, namely transport, inventory, motion, waiting, over-processing, overproduction, defects, and skills, *mura* refers to process variability, and *muri* to excessive workload. To this end, Lean principles are supported by a set of established tools that enable the operationalisation of the key goals (Varela et al., 2019). One of the most widespread Lean tools to uncover waste in production is Value Stream Mapping (Richter et al., 2023; Tortorella et al., 2020). In their renowned book, "Learning to See", Rother & Shook (2003) defined the key steps of Value Stream Analysis to enable the optimisation of production processes by improving the whole. These steps include value stream selection through product family identification, current-state mapping, future-state mapping, and implementation (Rother & Shook, 2003). After mapping the value stream corresponding to a product family, the material and information flows can be analysed and improved using various approaches, including pull production, production levelling, and line balancing.

Establishing pull is at the core of both Lean principles and value stream design principles (Rother & Shook, 2003; Womack & Jones, 1997). Pull production, together with levelled demand, enables a situation in which downstream processes obtain precisely the materials they need when they need it and upstream processes

are efficient (Smalley, 2004), thus limit the WIP (Hopp & Spearman, 2011). In line with the steps of value stream mapping defined by Rother & Shook (2003), creating a pull system can be achieved for one value stream at the time, focusing on a specific product family (Smalley, 2004).

Production levelling ("Heijunka") focusses on distributing the production volume and the production mix evenly over short periods. Thus, the levelling pattern is comprised of repetitive sequences of short periods that schedule the production according to fixed intervals. By decoupling production and customer demand, levelling enables the reduction of waste, unevenness, and overburden (i.e. muda, mura, muri) (Liker, 2020). Overall, a consistent or levelled production pace allows the production flow to predictable and enables quicker detection of anomalies (Rother & Shook, 2003). Conventional levelling approaches focus on manufacturing every variant in every interval period (i.e. "every part every interval" - EPEI), and they are suitable for large scale production with limited product diversity and stable demand (Deuse et al., 2013; Slomp et al., 2009). A solution for applying the levelling principles to high mix-low volume production environments is to create family-oriented levelling pattern (i.e. "every family every interval" - EFEI), which requires dividing the production mix into roughly equal-sized families (Bohnen et al., 2011, 2013).

In the context of the production line assembling of a product, balancing the volume of production per shift is an established challenge (Sivasankaran & Shahabudeen, 2014). While single-model assembly lines allow the production of a single product, mixed-model assembly lines allow the simultaneous assembly of different product variants or models (Sivasankaran & Shahabudeen, 2014). For mixed-model assembly line balancing problems, defining product families is an important step. By identifying a sufficiently large subset of products (i.e. product families), it is possible to reduce the idle times generated by the balancing lines that produce different variants. Then, the balancing problem can be solved for each product family before a setup-optimal sequence of product families is computed for a given interval (i.e. EFEI) (Deuse et al., 2013).

After establishing the relevance of value streams for the successful implementation of Lean production, the following section presents the existing body of knowledge on how value streams can be identified.

2.2 GROUP TECHNOLOGY FOR IDENTIFYING VALUE STREAMS

Identifying value streams, and in particular the associated material flows, requires the definition of product families and the identification of the related processing steps or resources. To this end, Group Technology is an established engineering philosophy that enables the identification of homogenous products from a manufacturing point of view and the processes required to make them (Burbidge, 1989; Hameri, 2011). For Industrial Engineering, Group Technology is essential for production optimisation (Deuse et al., 2022). The term Group Technology was first introduced by Mitrofanov as part of his research on the relationship between product shapes and processing methods. Central to his research was the idea that a lathe could make a "group" of similar parts (Mitrofanov, 1961). Then, the term evolved from group of parts to set of machines when an engineering company expanded Mitrofanov's initial idea by adding additional machines to form a "group" that completes all parts it made (Burbidge, 1991). Eventually, Burbidge introduced Production Flow Analysis as a method for planning Group Technology (Burbidge, 1991). Thus, machines are grouped according to the routing of a family of parts (Burbidge, 1989), resulting in a factory organisation in which units (groups) complete all parts they make (Burbidge, 1989, 1992). A detailed overview of Production Flow Analysis and its evolution is presented in the chapter dedicated to the systematic literature review (i.e. chapter 3).

The underlaying philosophy of Group Technology is transferring the advantages of mass production to job-shop production. While traditionally manufacturing was organised according to processes (i.e. process organisation) in which organisational units were specialised in specific manufacturing processes, product organisation enabled by Group Technology allows to achieve several advantages (Burbidge, 1991). These include lead time reduction, reduction in variation, better quality, bottleneck reduction, simpler production planning, and low inventories (Hameri, 2011).

In recent years, as industrial processes become more complex, simplification becomes more difficult, and searching and finding similarities in industrial processes is increasingly challenging. Thus, the traditional procedures underlaying Mitrofanov's or Burbidge's methods for Group Technology reach their limits.

However, the capabilities introduced by Industry 4.0 have unlocked the potential to couple these established methods with new technologies to improve their effectiveness. As such, combining the principles of Production Flow Analysis with the consistent use of production data enables streamlining the process of searching for similarities in highly complex systems (Deuse et al., 2022).

As a result of the new opportunities provided by Industry 4.0, recent research has focused on combining established IE methods and principles with the capabilities of new digital technologies. While there is extensive research examining the synergies between Lean production tools and Industry 4.0, PFA has received limited attention in existing literature. Overall, the principles of Lean production have been adopted by companies across the world because of their ability to increase productivity, customer satisfaction, and profitability (Rosin et al., 2020). As a result, recent research has investigated the synergies between Lean tools, that are mostly free of information technology, and Industry 4.0 technologies that enable inter-connectivity and growing decision-making capabilities of systems. A mapping study (Rosin et al., 2020) reviewing contributions that combine Industry 4.0 with Lean production principles and techniques, has observed that technologies including simulation, IoT, cloud computing, augmented reality and big Data Analytics have been exploited to improve the effectiveness of Lean principles such as takt time planning, pull systems, Jidoka and waste reduction. Therefore, companies should continue applying Lean methods and tools while improving their effectiveness using Industry 4.0 technologies (Rosin et al., 2020). In line with this conclusion, this research argues that in modern complex environments the integration of Industry 4.0 technologies, particularly Process Mining and Data Analytics, with PFA principles enables effective value stream identification, which is an essential step for the application of Production Planning and Control techniques, including Lean methods.

2.3 INDUSTRY 4.0 CAPABILITIES AND TECHNOLOGIES FOR IMPROVING THE EFFECTIVENESS OF VALUE STREAM IDENTIFICATION

Industry 4.0 has emerged as new technological paradigm shift promoting the adoption of new digital technologies in manufacturing systems. The concept was introduced at the Hannover Fair in 2011 as part of an initiative to improve German competitiveness in manufacturing (Kagermann et al., 2013). As a result of the

integration of physical and digital systems, Industry 4.0 changes the traditional trade-offs among priorities of cost, flexibility, speed, and quality (Olsen & Tomlin, 2020), and unlocks new opportunities for increased efficiency of processes and product differentiation. To harness these growth opportunities, companies are required to upgrade their digital competencies and capabilities (Schuh et al., 2020).

According to an Industry 4.0 maturity model (Schuh et al., 2020), companies can upgrade their systems through six stages, with each building on the previous one and each enabling new capabilities. The two initial stages, computerisation and connectivity, are the basic requirements for digitalisation. While *computerisation* represents an isolated use of information technologies, *connectivity* refers to the interconnected but not yet completely integrated use of IT systems (Schuh et al., 2020; Zeller et al., 2018). The four following stages represent the capabilities required for Industry 4.0.

Visibility enables process recording from start to finish through the use of technologies such as sensors, which results in up-to-date models of factories (Schuh et al., 2020). These models, also called Digital Shadows, are data profiles containing the information of a system's characteristics and historical, current, and future status. By providing an overview of what is happening on the factory floor at all times through integrated data models, Digital Shadows enable data-driven decisions (Tao et al., 2019). Despite of these benefits, generating Digital Shadows is not trivial, as companies face challenges including data being located across different sources (i.e. decentralised silos) and insufficient data collection (Schuh et al., 2020). Another challenge is represented by the fact that even when data is available, often companies do not know how to exploit the data they have effectively (Kusiak, 2017). To enable visibility, it is necessary to always create up-to-date models of the entire factory, instead of focusing on specific analyses (Schuh et al., 2020).

Transparency enables the causal understanding of events and correlations through the analysis of data in engineering contexts. Based on the Digital Shadows generated by the previous stage, data is analysed by applying engineering knowledge to understand why something is happening, ultimately enhancing process knowledge and support complex decision-making (Schuh et al., 2020; Zeller et al., 2018).

Predictive capacity enables the analysis of future state scenarios through simulation to anticipate future developments. The Digital Shadow generated by the previous

stages is used to project and analyse various scenarios and assessed in terms of likelihood (Schuh et al., 2020; Zeller et al., 2018).

Finally, *adaptability* takes predictive capacity to a higher level by enabling automated actions and automated decision-making (Schuh et al., 2020). By using data and information produced by the Digital Shadow, it is possible to automatically trigger corrective actions without human intervention, thus operating as a Digital Twin. Although conceptually similar to Digital Shadows, Digital Twins are superior as they provide more accurate and comprehensive knowledge. By merging data from both the physical and virtual environments, Digital Twins generate high-fidelity models operating together with the system, and they are able to provide comparisons between real and simulated performances while highlighting deviations (Tao et al., 2019).

The realisation of the Industry 4.0 capabilities mentioned above strongly depends on the inter-connection and computerisation of traditional manufacturing environments (Y. Lu, 2017). Relevant enabling technologies include Internet of Things (IoT), Cloud computing, and Cyber Physical Systems (CPS).

IoT is defined as "a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual 'Things' have identities, physical attributes, and virtual personalities and use intelligent interfaces, and are seamlessly integrated into the information network" (S. Li et al., 2015, p. 244) . Initially referring to the uniquely identifiable interoperable connected objects using RFID, IoT technology is now used in combination with other technologies such as sensors, actuators, and mobile devices (L. Da Xu et al., 2018). As a result of integrating tags and sensors into "things", IoT enables information gathering, storing and transmitting. For manufacturing processes, this means that identification technologies allow to track and monitor products along their life cycle (S. Li et al., 2015).

Cloud computing enables processing of large amounts of data and supports intensive computation. As opposed to traditional manufacturing environments where computing resources (e.g. servers, databases) are separate, cloud computing represents a solution for centralised computation and storage that supports complex decision-making tasks (L. Da Xu et al., 2018).

In Cyber Physical Systems (CPS), physical and software components are completely integrated, thus merging the physical and virtual world. As such, it is considered a key enabling technology for Industry 4.0, where multiple CPS form the Cyber-Physical Production System (CPPS) for which equipment becomes increasingly intelligent. Together with IoT technologies, CPS enable the development of smart factories, that are at the core of Industry 4.0 (Alcácer & Cruz-Machado, 2019; L. Da Xu et al., 2018).

As information technologies are integrated into manufacturing processes, large amounts of data become available. In general, Industry 4.0 architectures are comprised of various layers, with the lower ones collecting and monitoring equipment on the factory floor, and the higher ones doing data analysis for decision-making purposes. This enables the generation of manufacturing data across various stages. Typically, data is collected from sources including equipment, products, and human operators by means of the IoT, RFID, and other sensors. This data is integrated and stored securely in warehouse systems or Cloud systems, and then processed to remove redundant or inconsistent information. Through Data Analytics, it is possible to generate valuable knowledge that can inform decisions about whether, when, and how to adjust the manufacturing processes and equipment (Tao et al., 2018).

The use of Data Analytics in industrial applications is also referred to as Industrial Data Science (Deuse et al., 2022). More generally, Artificial Intelligence (AI) has been attracting increasing interest among researchers and industry practitioners for its capability of automated knowledge acquisition. An important subset of AI is Machine Learning, which applies algorithms to structured data in order to learn from it and build models without being programmed manually (Stanescu et al., 2018). Another important research area closely related to Machine Learning is Data Mining. While Machine Learning focuses on the automated induction of models, Data Mining focuses on knowledge discovery through the extraction of meaningful patterns from empirical data (Hüllermeier, 2011). Within Data Mining, Process Mining has been increasingly used for data analysis in production as it is able to analyse event logs from the factory and discover, analyse and ultimately suggest improvements for existing manufacturing processes (Knoll et al., 2019).

Process Mining is an emerging area connecting data science and processes science. The first significant work in this field was done by Wil van der Aalst in 1998, when he developed the first algorithms able to discover process models in the form of Petri Nets from event data. He is considered the "Godfather of Process Mining" and over the years he has earned multiple awards, including the prestigious German award Alexander-von-Humboldt Professorship. Building on Wil van der Aalst's invention, an increasing number of use cases was discovered. Starting from the 2000s, Process Mining gained increasing interest within academia, while starting from 2010 it became increasingly adopted by companies including BMW and Siemens. In 2010, Wil van der Aalst published the first book on Process Mining, which got updated and extended in 2016 (i.e. "Process Mining: Data Science in Action") and became the main reference in the field (Reinkemeyer, 2024).

The increasing interest in Process Mining is due to its ability to facilitate the provision of insights based on factual data, enhancing process understanding and enabling improvements (van der Aalst, 2016). As opposed to Data Mining, which is a data-centric discipline focused on finding relationships and patterns in large datasets, Process Mining maintains a process-centric focus, providing as main output process representations in the forms of Petri Nets or BPMN (van der Aalst, 2016). Some of the questions that Process Mining can answer are: "What happened in the past?", "What is likely to happen in the future?", "How to control a process better?", "How to redesign a process to improve its performance?" (van der Aalst, 2016).

Similarl to Data Mining, Process Mining is a discipline based on data and its main data input is event logs. As a result, Process Mining algorithms for process discovery are able to transform the information contained in event logs into process models (van der Aalst, 2016). Event logs are sequential records of events where each event represents a specific step in the process (i.e. activities) and is related to a particular case. For each case a particular trace, that is a specific sequence of events can be recorded. In this perspective an event log can be seen as a collection of traces (van der Aalst, 2016).

The data used for Process Mining applications follows a life cycle similar to that of a typical Industry 4.0 architecture (Tao et al., 2018). Thus, raw data can originate from different sources scattered across the organisation. Information is recorded by physical devices (e.g. RFID, IoT), web services, ERP systems, and transportation

systems (Chandra Bose et al., 2012). Then, data needs to be extracted, transformed, and loaded (ETL) into a target system such as a data warehouse or a relational database that unifies information. Due to the big amounts of data store in the data warehouse, different event logs may be extracted. A coarse-grained scope is defined based on the use case and questions to be answered. Finally, the event log can be further filtered (fine-grained scoping) based on the initial results (van der Aalst, 2016).

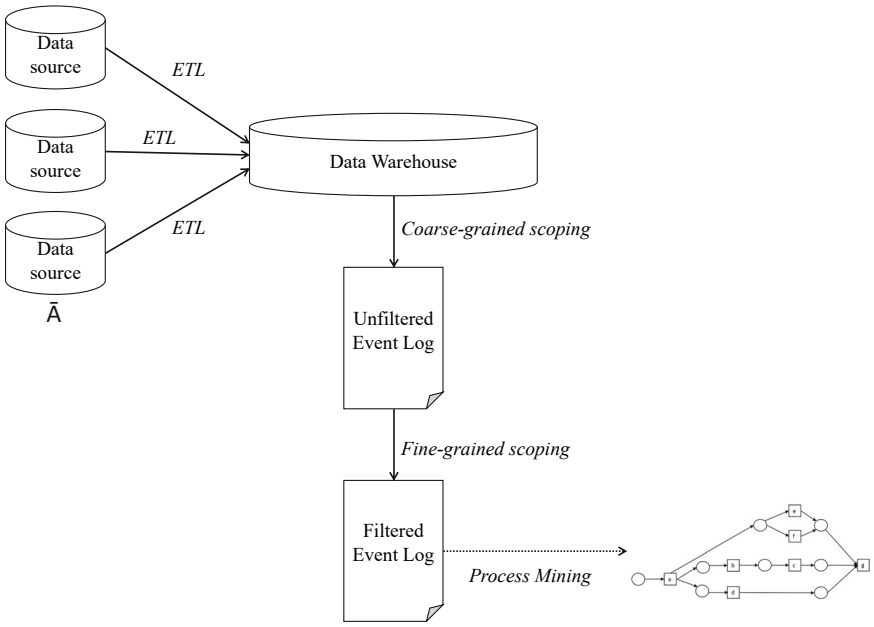


Figure 2-2: Process Mining workflow (adapted from van der Aalst (2016))

This process leads to the generation of an up-to-date digital model of factories represented using Petri Nets, process trees, or BPMN models. In other words, Process Mining creates a Digital Shadow of manufacturing processes and unlocks visibility in production (van der Aalst et al., 2021). As enabling transparency requires the application of the engineering knowledge to the captured data (Schuh et al., 2020), this research integrates the capabilities of Process Mining with the principles of PFA to enable the identification of key value streams and ultimately support the application of essential Process Planning and Control methods in complex environments. As visibility and transparency represent the groundwork for unlocking advanced Industry 4.0 capabilities (Schuh et al., 2020), this work also provides the foundations for predictive and adaptive capabilities.

The following chapter systematically explores existing research in Process Mining and PFA, and it provides a detailed overview of the areas for future developments as well as the research gaps.

Chapter 3: Systematic Literature Review

This chapter is published in the following conference proceedings:

Tomidei, L., Sick, N., Deuse, J., & Clemon, L. (2022). Production Flow Analysis in the Era of Industry 4.0 : How Digital Technologies can Support Decision-Making in the Factory of the Future. In 2022 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 1-15). Piscataway, USA: IEEE. doi:10.23919/picmet53225.2022.9882711

This section aims to clarify the research gaps and strengthen the validity of the research question. Therefore, the goal of this study is twofold (1) understanding the evolution of intelligent approaches in relation to Production Flow Analysis (PFA), and (2) capturing the potential of Process Mining to support decisions in manufacturing environments. By comparing the developments in PFA with the capabilities of Process Mining, it is possible to identify potential synergies and draw a future research agenda. The goal is to provide contribution to both academia and industry practice, with a particular focus on enhancing decision-making in manufacturing practice.

3.1 METHODOLOGY

A systematic literature allows to explore the existing research by using a structured approach. The main steps employed for this study follow the guidelines provided by (Levy & Ellis, 2006) and include selecting the database, defining the search strings and related filters as well as listing the inclusion and exclusion criteria used to select the final pool of relevant documents.

The database used to review a broad range of high-quality scholarly literature is Scopus. In accordance with the twofold purpose of this study, two research bodies have been analysed. Table 3-1 summarises the search strings and filters. The keywords selected are based on the definitions provided in the previous section. In particular, the first search string combines definitions related to PFA with Industry 4.0, AI and the related subsets. This ensures that all types of intelligent approaches are returned by the search.

Table 3-1: Search string

Research area	Search string	Filters
Intelligent approaches for PFA	TITLE-ABS-KEY ((("cellular manufacturing" OR "cell formation" OR "cell design" OR "group technology" OR "Production Flow Analysis") AND ("industry 4.0" OR "industrie 4.0" OR "artificial intelligence" OR "machine learning" OR "data mining" OR "Process Mining" OR "neural networks")))	Document types: articles, reviews
Process Mining approaches in manufacturing	TITLE-ABS-KEY ("Process Mining" AND manufacturing)	Document types: articles, reviews

Table 3-2 lists inclusion and exclusion criteria for each research area. Based on the definition of Production Flow Analysis proposed by (Burbidge, 1989), only Group Technology is considered for the cell formation problem, while other aspects such as layout design and machine scheduling are excluded, as they do not strictly focus on the concept of Group Technology (i.e. forming families of parts and machines). Production Flow Analysis focuses on identifying families of parts that are processed by the same machines. Therefore, relevant papers are considered those that focus on Group Technology or the cell formation problem and use intelligent approaches. For example, Singgih (2021) proposes a Machine Learning based framework to identify factors that affect the system throughput level in a semiconductor fab. Although the approach is named Production Flow Analysis, the aims do not include those proposed (Burbidge, 1989) (i.e. part-machine grouping) and therefore the study has not been included.

Table 3-2: Inclusion Criteria

Research area	**Inclusion criteria**
Intelligent approaches for PFA	The methodology fits in the definition of Production Flow Analysis proposed by Burbidge. The paper addresses group technology and the cell formation problem – either in the form of

	application or literature review (scheduling aspeects and layout design are not considered). The cell formation problem is solved through an intelligent approach.
Process Mining approaches in manufacturing	The paper applies Process Mining techniques to a manufacturing context. The paper applies Process Mining to a context applicable to manufacturing.

In order to define intelligent approaches, the framework provided by Burggraf et al. (2021) has been applied. Although traditionally within intelligent resolution approaches of most interest fuzzy systems and expert systems are included, the authors exclude these from their framework. In line with other authors (Drira et al., 2007; Hosseini-Nasab et al., 2018) fuzzy data and fuzzy systems are classified as data types instead of resolution approaches. As expert systems are a representation of an expert's knowledge, they lack the capability of learning independently and being able to find unseen correlations and solutions, which Machine Learning systems have. Therefore, within intelligent approaches, their framework focuses on Machine Learning approaches and the different models they are comprised of. In particular, Artificial Neural Networks (ANN) have been attracting most research interest and for that reason this keyword has been added to the search string (Burggraf et al., 2021). In addition to the elements included in the framework proposed by Burggraf et al. (2021), data mining and Process Mining have also been included in the search string as they both represent sub-domains of artificial intelligence.

For the review of Process Mining applications in manufacturing contexts, relevant papers describe applications in manufacturing contexts or applications that are applicable to manufacturing contexts.

3.2 RESULTS

The final samples have significantly different sizes. The literature focusing on intelligent approaches for PFA includes 122 papers, while the one on Process Mining approaches in manufacturing includes 49 papers. The main difference lays in the time when publication activity started for the two research areas. By comparing the number of relevant papers per year (see Figure 3-1), it is possible to see how the research

interest varies over time. Machine learning and data mining techniques have been used by academics and industry practitioners for decades. In fact, soon after Production Flow Analysis was proposed by Burbidge in 1970s, solutions of intelligent approaches for group technology have been steadily explored. In line with the findings provided by another recent literature review, results show that this research area experienced its maturity before the 2000s and maintained popularity in the following years (YounesSinaki et al., 2023). On the other hand, research on Process Mining applications in manufacturing contexts has gained interest only in the last decade, despite of the fact that data mining tools have been available for much longer. The reason is that Process Mining provides a representation of production processes based on event logs collected from the factory, and through the advent of Industry 4.0, manufacturing systems have become more digitalised, and it has been possible to collect an increasing amount of data. Therefore, studies in this field have been exploring how Process Mining techniques can assist informed decision-making for manufacturing systems.

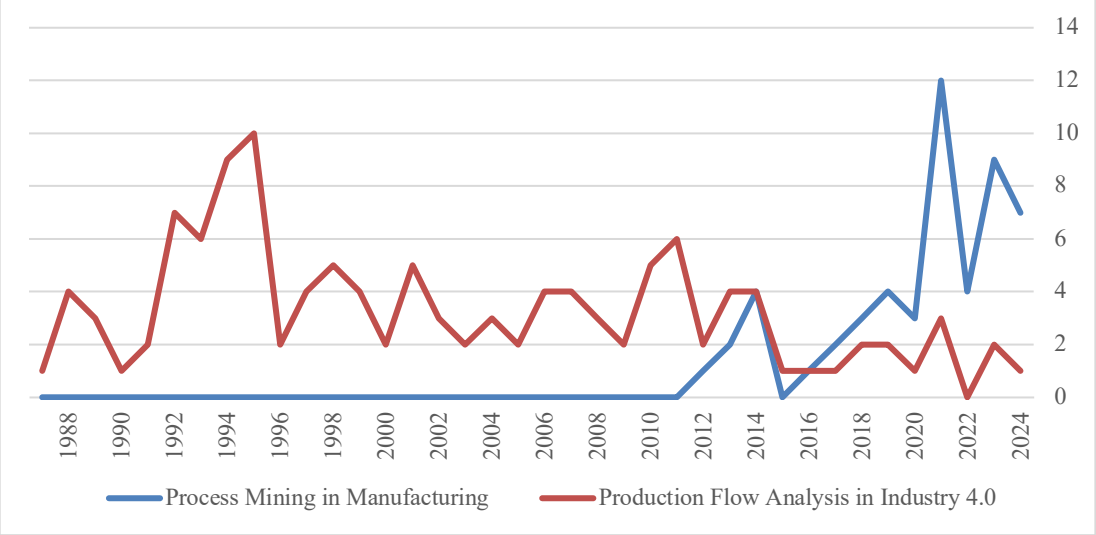


Figure 3-1: Research contributions in Process Mining and Production Flow Analysis over time

As expected, most contributions for both research areas focus on engineering and computer science, respectively. Interestingly, business, management and accounting, and decision sciences follow next for both research streams, representing 8.6% and 9.1% of contributions for PFA, and 8.7% and 7.9% for Process Mining. This highlights the importance of these techniques to support decision making processes and business decisions.

By examining results more in depth, regardless of the consistent number of papers resulted from the literature review for Production Flow Analysis, there is only a few that include industrial applications. While the cell formation problem is solved through intelligent approaches, most of these applications use traditional methodologies for the collection of process features and they manually extract data from the field as well as from other sources and databases. Production Flow Analysis can be supported by several techniques, most of which use Machine Learning approaches. For these applications, technologies are used to support the resolution phase as they exploit their computational capabilities to provide optimal solutions. Therefore, most of the studies in the literature illustrate algorithms that are validated through experiments. Only few of these validate their approach through the application to an industrial case, and yet production data is collected manually. On the other hand, studies related to Process Mining techniques in manufacturing contexts tend to provide substantially more case studies, and they focus on applications which are built upon real production data that is extracted automatically from event logs. Figure 3-2 illustrates the portion of case studies found in the two research areas under examination.

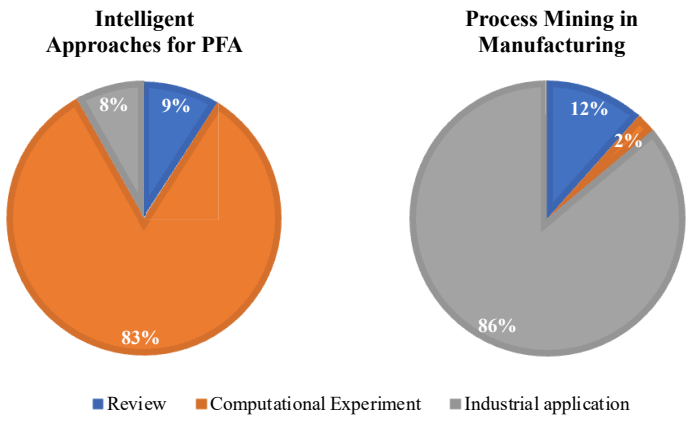


Figure 3-2: Comparison between research areas

3.3 DISCUSSION OF RESULTS

3.3.1 Intelligent Approaches for Production Flow Analysis

Intelligent approaches to solve the cell formation problem have been studied and adopted since 1980s. For that reason, the current literature presents a large number of studies and several reviews. In this portion of the literature, Machine Learning techniques are often used in combination with more traditional heuristic approaches in

order to find an optimal solution for Group Technology. Experiments are carried out to validate performance and only few studies illustrate applications involving industrial case studies. Yet, for most of those production data is collected manually, and then fed into a proposed algorithm. Figure 3-3 illustrates the different approaches that emerged from the literature review. The majority of them relies on Machine Learning techniques, often combined with metaheuristics. Appendix A provides a more specific analysis of the approaches emerged from this part of the literature.

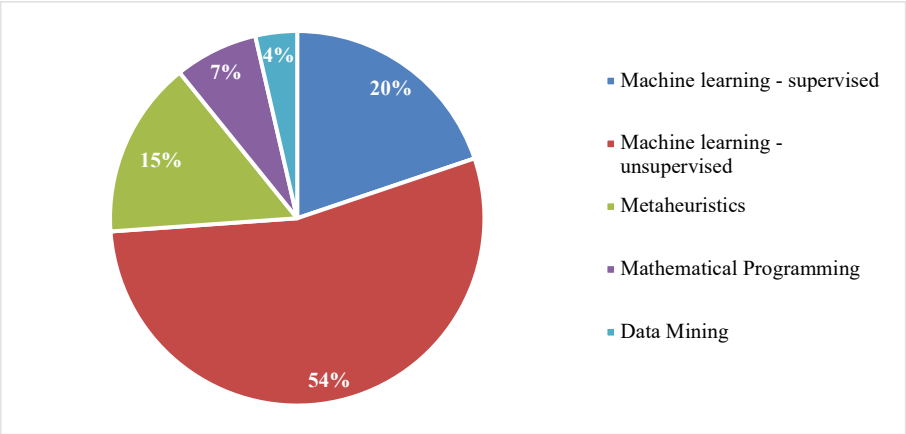


Figure 3-3: Techniques for Intelligent Production Flow Analysis

Machine Learning techniques include a variety of algorithms, which can be classified as either supervised learning or unsupervised learning. In general, supervised learning algorithms are able to gain knowledge from a dataset in which targets are labelled cases, and then use these to predict new cases (unlabelled). Unsupervised learning algorithms handle datasets that have not been classified and group them into clusters (Berry et al., 2019).

In the last few decades, the interest in Machine Learning algorithms for Group Technology applications has varied. Figure 3-4 illustrates the different research trends that have formed throughout the years. It is possible to note that the research interest in specific intelligent approaches has not been constant. Most notably, while supervised ANN, ART1 neural networks, and self-organizing maps have been extensively explored between 1990 and 2000, the research interest in the following decades has visibly dropped. On the other hand, metaheuristics approaches such as ant colony and genetic algorithm have regained popularity in the last decade.

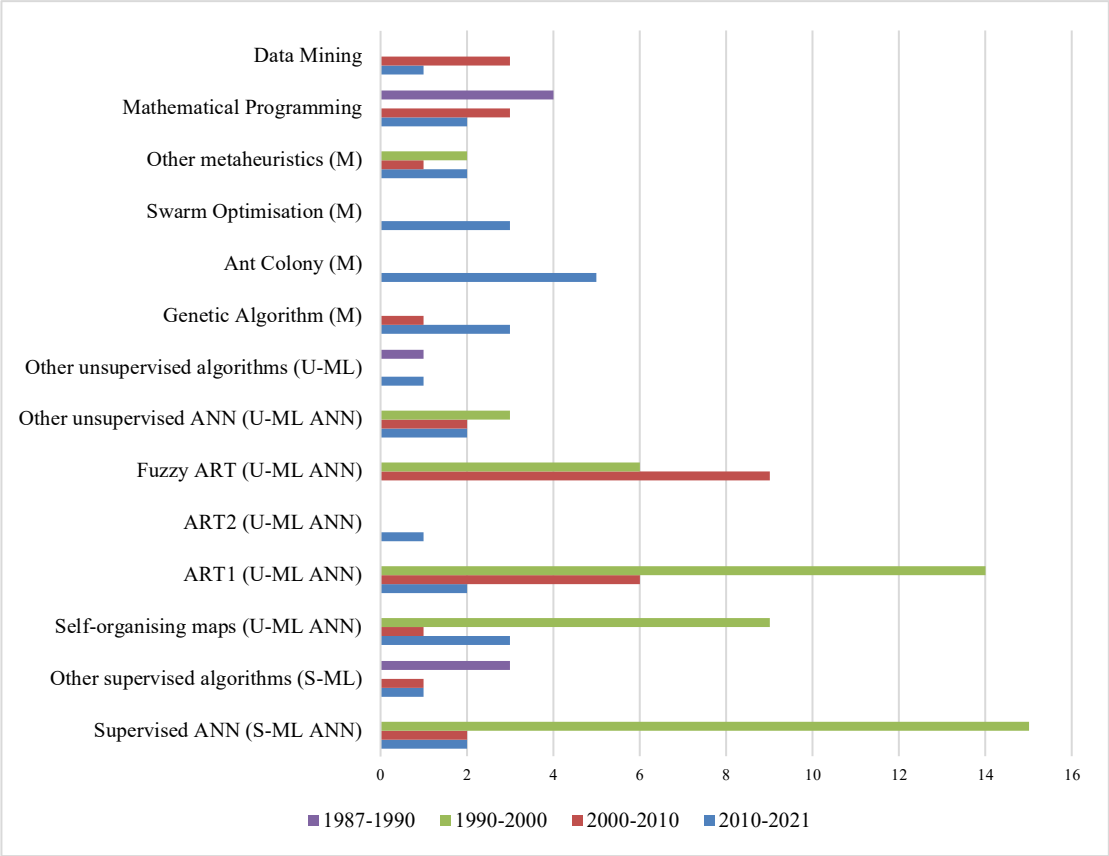


Figure 3-4: Research trends of intelligent approaches for PFA over time (M: Metaheuristics, ML: Machine Learning, S: Supervised Learning, U: Unsupervised Learning)

The literature review also returned a considerable number of reviews, demonstrating a generally high interest in the topic. YounesSinaki et al. (2023) investigate the cellular manufacturing design problems trends between 1996 and 2021 by providing a classification for numerous publications concerning the problem formulation and solution procedures. Ghosh et al. (2014) conducted a review on the evolvement of intelligent approaches based on artificial neural networks (ANN) in the context of cellular manufacturing. Similarly, El-Kebbe & Danne (2006) proposed an overview of neural networks-based approaches for machine-part grouping. Chattopadhyay et al., (2013) slightly broadened the scope by proposing a review on ANN and genetic algorithms approaches. Papaioannou & Wilson (2010) conducted a review of the evolution of all resolution approaches for the cell formation problem. A decade earlier, Venugopal (1999) conducted a review on all soft-computing approaches used for the group technology problem. Renzi et al. (2014) focused on a wider range of cellular manufacturing problems, including cell formation problem,

layout and scheduling, and propose a review of intelligent approaches for these three domains.

As the focus of the reviews suggests, ANN is one of the most popular intelligent approaches used for cell formation problems. Table 3-3 shows the types of ANN that emerged from the literature review for this study.

Table 3-3: ANN Approaches for Production Flow Analysis

ML Approach	ANN Algorithm	Occurrence
Supervised	Back propagation	60%
	Forward propagation	27%
	Recurrent	13%
Unsupervised	ART1	36%
	Fuzzy ART	25%
	Self-organising maps	23%
	Interactive activation and competition	7%
	Competitive rule	3%
	Fuzzy self-organising maps	2%
	ART2	2%
	Fuzzy min-max	2%
	Transiently Chaotic	2%

Very few contributions in this research area propose methods used for industrial applications, and although these use intelligent approaches for the resolution of the cell formation problem, data is still collected using traditional methods. The most popular intelligent approach for industrial applications is ANN. This is often combined with other heuristic approaches and sometimes modified to handle fuzzy datasets that represent uncertainty. S. C. Y. Lu & Ham (1989) proposed a goal-directed part classification framework for group technology based on an unsupervised learning algorithm – conceptual clustering. The authors demonstrated their approach through a case study using John Deere's database which included over 1000 rotational parts. The variables representing part features were selected manually based on experts' knowledge and these were used as input for the clustering algorithm proposed by the authors. F. F. Chen & Sagi (1995a) developed a decision support system that assist

manufacturing cell design as well as definition of cell control functions. By using simulation technology and ANN, the authors propose a methodology that simultaneously focuses on the cell formation problem and the required cell control functions. The approach is applied to an industrial case study consisting of an automated manufacturing cell. As the methodology is based on an existing industrial scenario, the authors define as a first step the description of operations (DO) and the definition of functional specifications (FS). These parameters are collected manually at the start, and they are particularly important as the concurrent configuration of the cell and its control functions is based on them. Depending on the simulation output and the predicted unit cost, DO and FS may need to be revised. Seo & Park (2004) proposed a methodology to solve the recycling cell formation problem. In this scenario, group technology is used to classify products in recycling part families in their end-of-life phase. The authors developed a methodology based on a fuzzy C-mean algorithm combined with a modified fuzzy neural network. The approach has been applied to a scenario which describes the disposal of refrigerators. The dataset includes information collected by an experiment of disassembly disposal refrigerators and a recycling centre. This information needed to be reviewed manually to extract the relevant attributes. Based on these attributes, the methodology was able to define recycling cells. Mahmoodian et al. (2019) developed an intelligent particle swarm optimisation algorithm that exploits self-organising map neural networks to solve the cell formation problem. The authors validated their methodology by applying it to a company from the agricultural manufacturing sector. The approach produces an incidence matrix in a diagonal block form representing the optimal cell formations. For the industrial application, the incidence matrix has been first defined manually and then used as a base for the algorithm. Aloudat et al. (2008) built a tool that is exploits data mining techniques and ANN to examining factors that impact cell quality performance and suggest improvements. Recommendations focus on a variety of aspects including the formation of family products, part processing sequence and machine process capability. As data mining allows to discover meaningful correlations and patterns from structured datasets, the authors use these techniques to analyse how different variables in running family products interact and affect the cell quality performance. The approach has been applied to an industrial case study in which 261 quality reports were collected for validation.

3.3.2 Process Mining in Manufacturing

Process Mining is explored in a variety of ways in order to propose approaches that can be applied to existing industrial scenarios. The potential of this technique lays in its ability to generate automated process models, which allow to obtain a transparent and reliable representation of the shop floor. Based on this, it is possible to analyse and estimate the performance of the manufacturing system or more generally to support the application of other techniques such as value stream mapping.

Traditionally, Process Mining applications could be distinguished into three main categories, namely process discovery, process conformance, and process enhancement W. M. P. van der Aalst, 2012). More recently, van der Aalst (2022) presented an updated framework categorizing six types of Process Mining tasks, namely process discovery, conformance checking, performance analysis, comparative Process Mining, predictive Process Mining, and action-oriented Process Mining (van der Aalst, 2022).

Through process discovery, process models are generated based on the example behaviour contained in the event log. Conformance checking compares an event log (i.e. observed behaviour) and a process model (i.e. modelled behaviour) for evaluation purposes. Performance analysis aims to uncover problems such as limited productivity, excessive rework, and tardiness to improve processes. Comparative Process Mining uses as input multiple event logs referring to different locations, categories, or time periods to draw insights. Predictive Process Mining is used in combination with Machine Learning to create predictive models to foresee performance problems. Finally, action-oriented Process Mining aims to turn diagnostics into action by enabling an understanding of events occurred in the past, happening in the present, and likely to happen next in the process (van der Aalst, 2022).

The literature provides a range of studies within the manufacturing context, whose purpose can be framed mainly into three of the six categories defined by Van Der Aalst (2022), namely process discovery, performance analysis, and predictive Process Mining (see Figure 3-5).

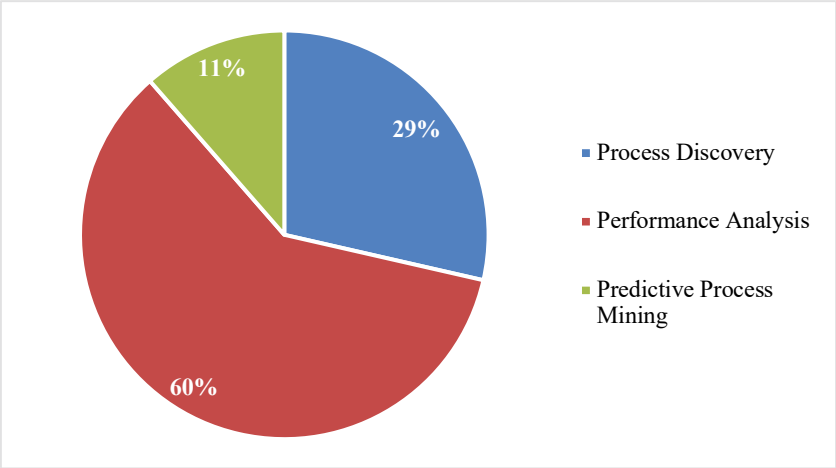


Figure 3-5: Process Mining types in Manufacturing

The majority of contributions use Process Mining to analyse the performance of manufacturing systems and assist domain experts in production planning and control decisions. This portion of the literature evaluates system performance by using Process Mining to execute specific operations management tasks, analyse specific KPIs, or improve the effectiveness of established Lean tools and techniques for production analysis and improvement.

Several studies focus on specific production tasks including bottleneck analysis (Kumbhar et al., 2023; Laghouag et al., 2024; Rudnitckaia et al., 2022), quality assurance (Cho, Park, Song, Lee, & Kum, 2021; Duong et al., 2021; C. K. H. Lee et al., 2014, 2016), and layout design and optimisation (Ceylan et al., 2023; Rismanchian & Lee, 2017). Other contributions focus on the analysis of specific parameters for performance evaluation including production throughput (Lugaresi & Matta, 2023) and waiting and execution time (J. Park et al., 2014). In some cases, the performance of a manufacturing system is evaluated using conformance analysis. For example, Lorenz et al. (2021) proposed a three-phase procedure in which two process models are built defining both the planned model and the real one, then they are compared to check process conformance and analyse production waste, and the findings are used to inform decisions for process improvement. A few studies use Process Mining to improve the effectiveness of Lean tools and techniques for process improvement. Horsthofer-Rauch et al. (2024) propose a framework for sustainability-integrated VSM using Process Mining, Tran et al. (2021) propose a method that integrates positional and manufacturing data to enable VSM and the collection of Lean KPIs, and

Knoll et al. (2019) used multidimensional Process Mining techniques to enable VSM and waste analysis for internal logistics.

Contributions using Process Mining for process discovery tasks are focussed on automating the generation of process models for creating Digital Twins in smart factories (Friederich et al., 2022; Yadav et al., 2023) or some of the complexities of manufacturing processes including assembly lines (Lugaresi & Matta, 2023), support processes (Lugaresi et al., 2023), and distributed departments (Sarno & Effendi, 2017). In general, real-life processes are often unstructured and complex, leading to the discovery of "spaghetti-like" process models (Kong et al., 2018). Based on this observation, two contributions have addressed the cell formation problem using Process Mining combined with the principles of Group Technology, that enable the definition of homogenous groups of resources (i.e. machines). Kong et al. (2018) developed a two-mode modularity clustering method based on new similarity measures to formulate a cell formation solution. The authors use an ordinal part-machine matrix that represents information about incidence and transitions, and they use an event log from a Dutch financial institute to validate their method through ProM software package. Delcoucq et al. (2023) propose a hierarchical cell formation algorithm to cluster resources that share the same behaviour. The approach aims to connect the resource perspective and activities perspective in Process Mining by using a normalised frequency-based incidence matrix combined with a Direct Clustering Algorithm, and an evaluation is made by measuring precision and recall values for the resources. Finally, although the study proposed by Rismanchian & Lee (2017) focuses on a healthcare context, the idea behind the methodology can also apply to a manufacturing context. In fact, the authors employed Process Mining to derive process models of noncritical patients and critical patients in the emergency department of an hospital. These are used to determine the optimal layout of the department with the goal of minimising the distance travelled by patients, maximising specific design conditions, and minimising the relocation costs. Similarly, in a manufacturing context, the same approach can be adopted to use Process Mining in order to assist layout design and analysis based on production data.

Existing contributions using predictive Process Mining focus on the estimation of specific process parameters depending on the use case. Thus, contributions focussing on sustainable manufacturing propose methods for estimating carbon

emissions (Wu et al., 2024) or for the simulation of energy consumption (Kaniappan Chinnathai & Alkan, 2023), while those focussing on scheduling and preventive maintenance planning focus on time estimation (Ruschel et al., 2021).

A synoptic table listing contributions of Process Mining in manufacturing is included in Appendix B.

3.4 SUMMARY

The research areas examined in this literature reviews share some similarities and demonstrate potential synergies. Both the research areas use intelligent techniques to address challenges emerging from manufacturing systems. In the context of Group Technology, the use of Machine Learning has been widely investigated and tested. However, as the literature review has shown, the majority of studies share similar characteristics: (a) they use intelligent approaches only for the resolution of the cell formation problem, (b) few studies validate their methodology through a case study, and (c) in these cases production data is collected manually. On the other hand, in the era of Industry 4.0, there is an increasing amount of data generated by manufacturing systems, which unlocks new opportunities for capturing the real production flows and gaining knowledge of the production floor at any time (Lorenz et al., 2021; Lugaresi & Matta, 2021). In this context, Process Mining has been receiving increasing interest among researchers for its ability to provide an accurate representation of production processes, which is crucial for managing the manufacturing system effectively. In fact, valuable decisions can only be taken when they are based on the assumption that process models are sufficiently aligned with the real system (Lugaresi & Matta, 2021). Accordingly, the literature review has demonstrated that Process Mining contributions include a significantly higher number of industrial applications.

3.5 RESEARCH IMPLICATIONS

Nowadays, manufacturing systems are extremely complex and having an accurate representation of real production processes is essential in order to build any other analysis. Research on Production Flow Analysis extensively validates the use of Machine Learning techniques for the effective resolution of Group Technology. Yet, studies in this area include scarce industrial applications and their main characteristics have not evolved consistently in the last decade. In order to provide contribution not only to academia but also to industry practitioners, new approaches should focus on

real production processes and accurate representations of existing manufacturing systems. Process Mining has been receiving increasing interest only recently, but it has already shown good potential for practical solutions. Contributions in this area address a variety of use cases, including Lean applications and layout design. Applications of Process Mining to Group Technology problems have hardly been explored yet, despite of the high potential for supporting decision making in this context. Existing contributions in the area maintain a focus on developing new clustering approaches and similarity measures (Kong et al., 2018), or clustering algorithms focused on the resource perspective of Process Mining (Delcoucq et al., 2023). Therefore, a future research agenda for Production Flow Analysis calls for approaches that demonstrate high practical applicability and exploit the data available on factory floors. In this context, Process Mining has the potential to assist the analysis and provide accurate models that can be used as a base for Production Flow Analysis.

Based on the detailed review provided in this study, the following suggestions have been developed:

- Advance approaches based real production data extracted directly from the factory floor to provide a reliable representation of production systems on which valuable decisions can be taken
- Apply the established principles of PFA to real production data and Process Mining techniques
- Further enhance the integration between digital technologies capturing the state of production systems at any point in time and decision-making models
- Provide companies with guidelines to implement such approaches

These recommendations aim to provide an opportunity to advance research in the established field of PFA and the emerging area of Industry 4.0.

It is important to note that this literature review has focused on Manufacturing settings and the potential synergies generated by PFA and Industry 4.0 technologies and Process Mining in particular. Some results (Rismanchian & Lee, 2017) have suggested that some other applicable fields may also contribute to this research. For example, Process Mining applications in the healthcare industry has the potential to be

applied and adapted to the needs of manufacturing contexts. Thus, expanding the research to such industries could enhance contribution. In addition to this, in order to specifically focus on PFA, the research design has excluded contributions that do not focus on the cell formation problem. These include studies on the scheduling problem within cells and layout design. However, these aspects are equally important for efficient operation of the factory. At the same time, as Process Mining allows to extract a structured dataset from the factory floor, this information could be used to develop advanced tools that address multiple aspects at the same time, including the cell formation problem, scheduling problem, and layout design.

After demonstrating the synergies between Production Flow Analysis and Process Mining, the following chapter provides a detailed analysis of the requirements for the application of Process Mining. This enables the definition of relevant use cases and directly informs the characteristics of the method for identifying key value streams.

Chapter 4: Industry Analysis and Requirements Definition

The quality of the result of Process Mining application is closely related to the input, that is event data (van der Aalst et al., 2012). Thus, data quality is crucial for the success of Process Mining (van der Aalst, 2016). In practice, the characteristics of data collection systems vary, meaning that the ability of generating quality event logs cannot be assumed consistent across the industry. Understanding the data requirements and the possible implications on the applicability of the method developed in this research enables the identification of suitable use cases.

In his work as inventor and pioneer of Process Mining, Wil van der Aalst has laid the essential groundwork for the definition of requirements in the field (Reinkemeyer, 2024). In particular, in establishing guiding principles for Process Mining, the "Process Mining Manifesto" published in 2012 by the IEEE Task Force on Process Mining and led by Wil van der Aalst (van der Aalst et al., 2012) sets key requirements for these applications.

4.1 DATA REQUIREMENTS

Production data including information about product routings is extracted in the form of event logs, that are a collection of events describing various process states. In regard to the definition of event logs, the IEEE Task Force on Process Mining states that: "All Process Mining techniques assume that it is possible to sequentially record events such that each event refers to an activity (i.e., a well-defined step in some process) and is related to a particular case (i.e., a process instance). Event logs may store additional information about events. In fact, whenever possible, Process Mining techniques use extra information such as the resource (i.e., person or device) executing or initiating the activity, the timestamp of the event, or data elements recorded with the event (e.g., the size of order)"(van der Aalst et al., 2012). In other words, event logs are generated based on the assumptions that a process includes different cases, events are associated to exactly one case, and they can have attributes associated (e.g. activity, time, cost, resources) (van der Aalst, 2016).

Definition: Events and Attributes (van der Aalst, 2016)

Let $\mathscr{E}$ be the event universe and AN a set of attribute names characterising events. For any event $e \in \mathscr{E}$ and name $n \in AN$, $\#_n(e)$ is the value of attribute n for event e. Standard attributes include $\#_{activity}(e)$ and $\#_{timestamp}(e)$.

Definition: Classifier (van der Aalst, 2016)

A classifier is a function that maps the attributes of an event onto a label used in the resulting process model, where $\underline{e}$ indicates the name of the event. If events are identified by their activity name, then $\underline{e} = \#_{activity}(e)$.

Definition: Cases, Traces, Event Log (van der Aalst, 2016)

Event logs consist of cases and cases consist of events. Let $\mathscr{S}$ being the case universe. Cases, like events have attributes. For any case $c \in \mathscr{S}$ and name $n \in AN$: $\#_c(e)$ is the value of attribute n for case c. Each case has a special mandatory attribute *trace*, $c\hat{} = \#_{trace}(e)$. A *trace* is a finite sequence of events $\alpha \in \mathscr{E}^*$ such event appear only once. An *event log* is a set of cases $L \subseteq \mathscr{E}$ such that each event appears at most one in the entire log. If an event log contains timestamps, then the ordering in a trace should respect these timestamps.

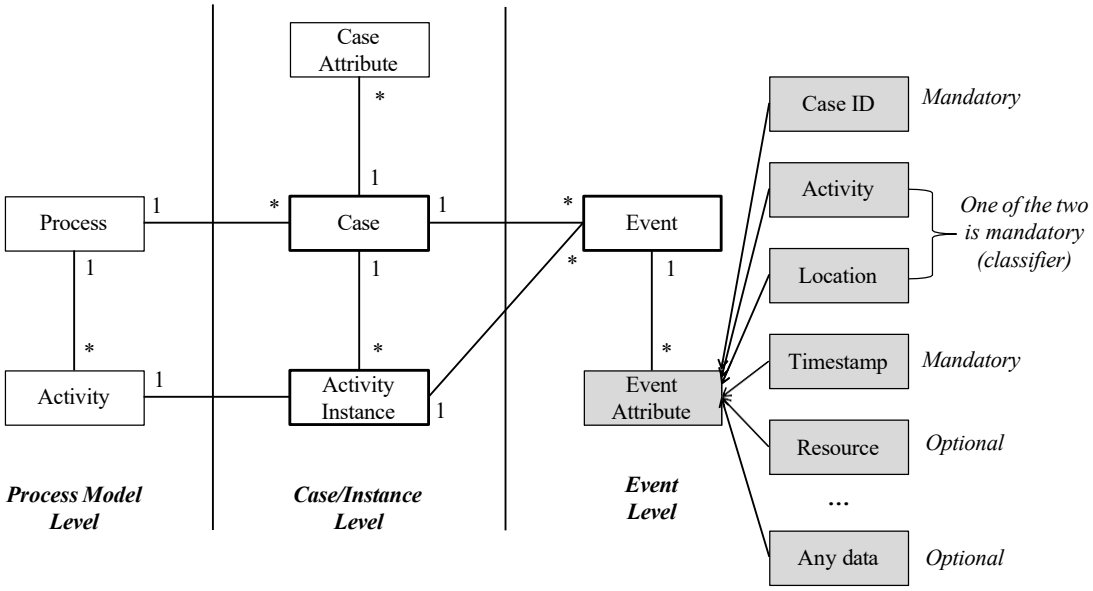


Figure 4-1: Class diagram and data quality (based on van der Aalst (2016))

In Manufacturing, event logs may contain information about a product, what scanning point it visited (where), and when, noting that multiple scanning points may be associated with different operations. Often event logs also contain additional information (i.e. attributes), specifying additional information such as order ID or batch number.

PRODUCT_NAME	SERIAL_NO	SCAN_POINT	OPERATION_ID	TIMESTAMP	…
A	XYZ	SC001	OP_1	29/11/22 10:05	…
A	XYZ	SC005	OP_2	29/11/22 10:23	…
A	XYZ	SC006	OP_3	29/11/22 11:21	…
..	..	..	..	..	…
D	XXX	SC012	OP_2	30/11/22 16:11	…
D	XXX	SC021	OP_3	30/11/22 17:22	…

Figure 4-2: Example of a complete event log (as used for this research)

Table 4-1: Event Log attributes description

	Attribute type	Description
Product Serial	Case ID	Unique product identifier (i.e. process instance)
Product ID	Case Attribute	Product name
Operation ID	Activity	Activity in the manufacturing process (may be associated to one or more scanning points)
Scanning Point	Location	Location where product get recorded
Timestamp	Timestamp	Time

The principles of Production Flow Analysis focus on the analysis of product routes for identifying product families and the resources they require (Burbidge, 1991). Defining product routings requires identifying individual products, resources, and times. Therefore, in order to use event logs for automated Production Flow Analysis applications, a minimum of three essential attributes need to be present in the event log, including unique product identifiers (i.e. case ID), information about work operations or location (i.e. Activity or Location), and timestamps. A case or process instance is associated to a record of n events, $e_1, e_2, \ldots e_n$. Thus, this process instance has a trace $\overline{e_1}, \overline{e_2}, \ldots \overline{e_n}$ associated to it (van der Aalst, 2016). By using activity names

as classifier, the trace corresponds to a sequence of activities. This enables the definition of operation sequences for each product in the factory. For PFA applications, it is essential to record product routings.

4.2 DATA COMPLETENESS AND USE CASES

Event logs should be complete, trustworthy (i.e. they reflect what happened in reality), have well-defined semantics, and the data should be safe (van der Aalst et al., 2012). However, in practice, even logs often have quality issues, including incompleteness, noise, and imprecision. Typically, possible issues regarding event log quality include missing data, incorrect data, imprecise data, and irrelevant data (Chandra Bose et al., 2012). Based on these considerations, Van Der Aalst et al. (2012) defined five maturity levels for event logs. Process Mining techniques can be used in scenarios corresponding to the top three levels of the maturity model. While theoretically Process Mining can be applied using event logs from the lower two levels, the analysis is problematic as the results are not considered trustworthy (van der Aalst et al., 2012).

Table 4-2: Maturity levels for event logs defined by van der Aalst et al. (2012) and implications for this research

Level	Characterisation	Implications for Production Flow Analysis applications
★ ★ ★ ★ ★	Event logs are of excellent quality, and they are generated automatically, systematically, and safely. They are trustworthy and complete, and privacy and security issues are considered. The events have clear semantics.	**Optimal Event Log** The event log does not need to be pre-processed. The event log can be directly used for PFA.
★ ★ ★ ★	Event logs are generated automatically, and they are trustworthy and complete. Notations are recorded explicitly (e.g. case or activity).	**Event Log Pre-Processing** The activity attribute (i.e. work operation) is not explicitly recorded. Instead, the event log only records the location products visit (i.e. scanning point).

		The case attribute specifying the association between product serial numbers and products (i.e. product ID) may also be missing.
⋆ ⋆ ⋆	Event logs are generated automatically but not systematically. They are trustworthy but not necessarily complete.	**Event Log Cleaning and Filtering** PFA can be applied. When possible, the event log needs to be repaired before being used for PFA.
⋆ ⋆	Information in the event logs is collected automatically but there is no systematic approach in deciding which events are recorded. Information may not be trustworthy.	Process Mining techniques cannot be applied. Therefore, it is not possible to have an automated approach for PFA.
⋆	Poor quality event logs that may not match physical processes (e.g. events manually recorded). Information may not be trustworthy.	Process Mining techniques cannot be applied. Therefore, it is not possible to have an automated approach for PFA.

Depending on the level of digital maturity of a manufacturing firm, not all the data points may be available and/or explicitly included in the event log. In order to develop an automated approach for Production Flow Analysis, it is necessary to take into account all the possible use cases (see Figure 4-3).

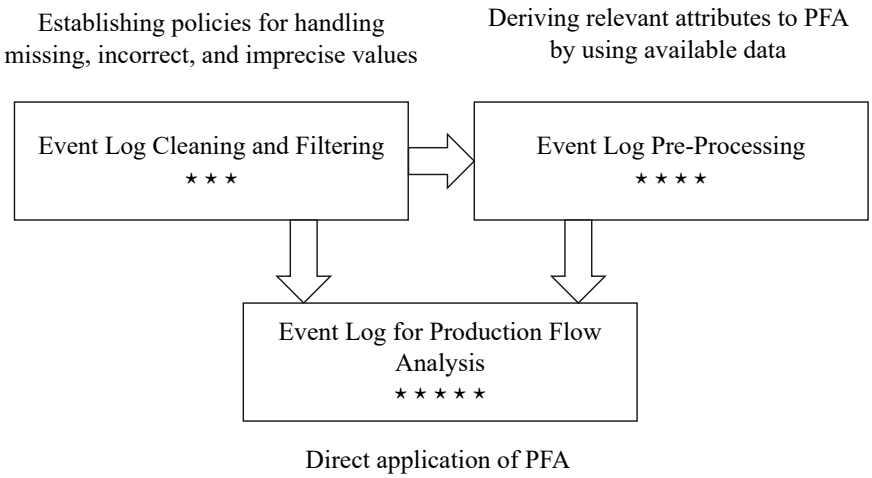


Figure 4-3: Conceptual structure for Process Mining-based PFA

4.2.1 Optimal Event Log

For excellent quality event logs, it is possible to proceed with the Process Mining-based method for PFA. This is valid whenever the event log contains the attributes presented in Table 4-1 and there are no missing, incorrect, or imprecise instances.

4.2.2 Event Log Pre-Processing

The simplest form of an event log would only include information related to product identifiers, the location in which they got recorded, and when (i.e. product serial number, scanning point ID, timestamp).

To allow a more detailed analysis, additional attributes may also be needed. On one hand, unique product identifiers (i.e. serial numbers) may be associated to a product name or ID. For example, if the event log was recording the manufacturing processes of an electronic goods manufacturer, the unique product identifier could represent a serial number and the product ID could represent the name of a product (e.g. phone model, tablet model). On the other hand, to identify the product routes, either a scanning point attribute or a work operation attribute may be present. While knowing the activity that products undergo enables direct discovery of their routing, knowing what scanning point they visited may require additional pre-processing steps before identifying value streams and generating the corresponding process models. In fact, multiple scanning points may be associated to the same work operation or activity. For example, this may be the case whenever there are machines executing the same activity in parallel and products can use any of the available machines.

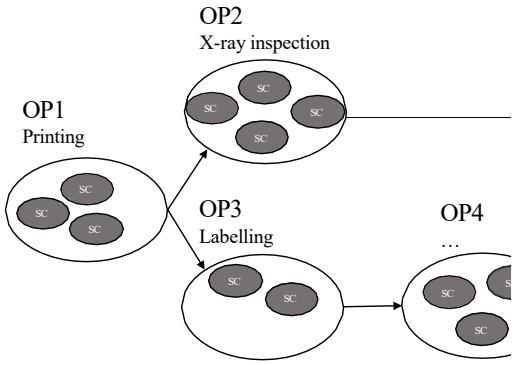


Figure 4-4: Visual representation of the relationship between operations and scanning points

4.2.3 Event Log Cleaning and Filtering

In practice, data quality problems can be common (Bose et al., 2013). Yet, the presence of these issues may not always be obvious. Missing data can be detected easily due to the absence of specific values for some of the events. In this case, the event log can be cleaned before continuing with any further data processing. Similarly, the limitations of information coarseness generated by imprecise attributes are also straight-forward. Instead, incorrect attributes may be harder to detect, especially if the behaviour is infrequent, that is attributes are occasionally recorded with uncertainty (Pegoraro & van der Aalst, 2019). Therefore, mechanisms for filtering out infrequent behaviour need to always be present.

Chapter 5: Method Development

Sections of this chapter have been published in the following conference proceedings:

Tomidei, L., Sick, N. & Mathieson, L. (2024). Data-Driven Value Stream Analysis Using Process Mining And Machine Learning. In 51st International Conference on Computers and Industrial Engineering (CIE51). Sydney, Australia.

Tomidei, L., Sick, N., Deuse, J. & Guertler, M. (2023). Extracting Key Value Streams using Process Mining and Machine Learning. In IEEE Conference on Engineering Informatics. 2023 IEEE Engineering Informatics, 1–7. https://doi.org/10.1109/IEEECONF58110.2023.10520644

This section outlines the overall methodology used to develop an automated approach for Production Flow Analysis. First, the core method is presented. This covers the optimal use case, in which a company is able to record event logs that are excellent quality and therefore no repairing or pre-processing is required. Second, the algorithm for event log pre-processing is presented. This is required for event logs that only record scanning points (i.e. the location that products visit) and not the associated operation. Finally, the implications for using event logs with process instances that are missing, incorrect, or imprecise are discussed.

An overview of the method for identifying value streams is illustrated in Figure 5-1.

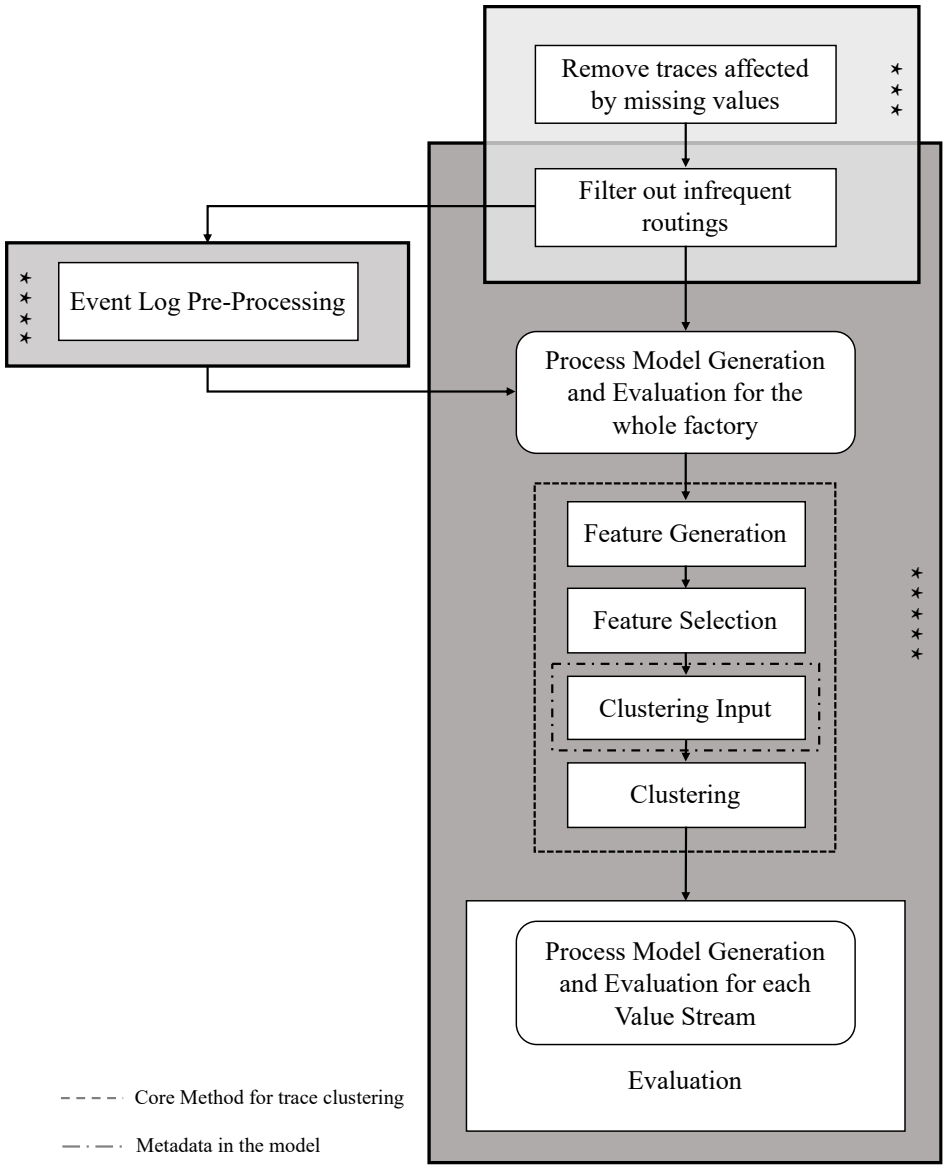


Figure 5-1: Overview of the method for value stream identification considering multiple use cases

5.1 OPTIMAL EVENT LOG

In the context of Industry 4.0, Process Mining has been increasingly used for data analysis in production as it is able to analyse event logs from the factory and discover, analyse and ultimately suggest improvements for existing manufacturing processes (Knoll et al., 2019). When the underlaying processes are complex, the corresponding even log is characterised by high levels of diversity, which results in "spaghetti models" and difficult interpretability (Song et al., 2009). This is the case for manufacturing systems that produce a large variety of products. Individual parts are associated to specific products, and their routings are recorded in an event log in the form of traces. In these scenarios, the assumption is that there are several structured process variants within one event log (Song et al., 2009). Based on the definition of value streams as sets of actions that bring a product family to the customer, it is possible to assume that these process variants represent value streams (i.e. product families and associated operations).

The identification of process variants as groups with homogenous traces is enabled by trace clustering, which is a common pre-processing technique in Process Mining. The idea behind clustering traces is to establish what makes two traces similar. This can be done by defining a profile for the trace, that describe its behaviour from a specific perspective (Song et al., 2009). For the purpose of this research, trace profiles describe product routings, and the trace clusters represent the key value streams.

The core method used for the identification of value streams from production data follows the framework for trace clustering in Process Mining proposed by (Zandkarimi et al., 2020), which is based on five consecutive steps, namely feature generation, feature transformation, feature selection, clustering input definition, clustering, and evaluation.

Thus, production data is extracted in the form of an event log and used to discover the process model representing all material flows. Then, the key value streams are identified using the key consecutive steps for trace clustering (Zandkarimi et al., 2020): feature generation, feature selection, clustering input definition, and clustering. The evaluation is done on both the process models of each value stream and clustering performance metrics.

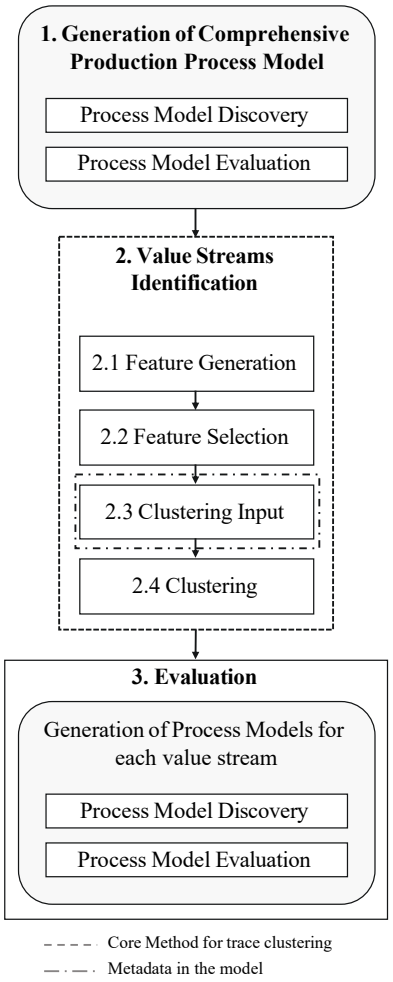


Figure 5-2: Method for value stream identification using optimal event logs

In production contexts, event logs are expected to include traces for each unique product moving across the factory floor. In order to be able to create the model of a production process, the event log should include information about unique product serials and associated product name or ID, the operations they go through, and the timestamps. As mentioned in chapter 4.2, operations may have more than one scanning point associated, meaning that multiple stations executing the same activity may be present.

PRODUCT_ID	SERIAL_NO	OPERATION_ID	SCAN_POINT	TIMESTAMP
P1	XYZ	OP_1	SC1	29/11/23 10:05
P1	XYZ	OP_2	SC3	29/11/23 10:23
P1	XYZ	OP_3	SC6	29/11/23 11:21
..	..	..	..	..
P5	XXX	OP_2	SC4	30/11/23 16:11
P5	XXX	OP_3	SC5	30/11/23 17:22

Figure 5-3: Example of an optimal event log

5.1.1 Process Model Generation

The event log representing the process flows in the factory is used to derive a process model representing all value streams. Typically, this results in a spaghetti diagram, that is unstructured and difficult to interpret and thus not suitable to inform decisions. However, this is a valuable starting point, as it allows to generate a big picture of product routings and all the operations involved.

Process Model Discovery

By definition, a "discovery algorithm is a function that maps an event log L onto a process model such that the model is representative for the behaviour of the event log". (van der Aalst, 2016, p. 163). Several process discovery algorithms have been introduced over the last two decades, and generally the three main approaches are alpha miner algorithms (van der Aalst et al., 2004), heuristic miner (Weijters & Ribeiro, 2011), and inductive miner (Leemans et al., 2014). The alpha miner algorithm was the first one to be introduced and it creates a process model based on the relationship between events algorithms (Peng et al., 2021). However, this type of algorithms has several shortcomings including underfitting, overfitting, and non-fitting (van der Aalst, 2016). Heuristic and inductive algorithms are considered advanced discovery techniques. While the heuristic miner is a frequency-based approach, the inductive miner is a divide and conquer strategy by splitting the event log into recursively sub-sections (Peng et al., 2021). In practical applications, it is essential to use discovery techniques that can handle noise and incompleteness issues, and inductive and heuristic miner algorithms have this ability (van der Aalst, 2016).

In this research, process models are generated primarily using the inductive miner discovery algorithm, as it represents a good solution when using large size event logs. It guarantees both soundness and re-discoverability and can be applied to a variety of use cases (Leemans et al., 2015). The inductive miner-infrequent is an extension of the inductive miner algorithm and enhances the original method by filtering infrequent behaviour. Compared to other algorithms, the inductive miner-infrequent generates models with lower fitness, higher precision, and equal generalisation and comparable simplicity (Leemans et al., 2014). Process models generated by the heuristic miner algorithm are included in Appendix C.

Process Model Evaluation

In order to evaluate the quality of a process model automatically discovered from an event log, conformance checking techniques are required (Adriansyah et al., 2015). These enable the comparison between the behaviour observed in the data and the one in the process model and the generation of diagnostics information (Berti & Van Der Aalst, 2019). The four main evaluation metrics are replay fitness, precision, generalisation, and simplicity.

Replay fitness indicates the extent to which the model can reproduce or replay the behaviour recorded in the event log (Buijs et al., 2012). Replay approaches can be alignment-based or token-based. Although alignment-based approaches return optimal results, their performance of complex models or large datasets is poor. Instead, token-based approaches are much faster and scalable. The application uses a trace of an event log and a Petri Net to measure the transitions that are enabled during the replay and whether there are remaining or missing tokens (Berti & Van Der Aalst, 2019).

Precision aims to check whether the process model is not underfitting the log by quantifying the extent of the behaviour allowed by the process model that is not observed in the event log (Buijs et al., 2014). Precision is measured by quantifying instances where the model deviates from the event log (Muñoz-Gama & Carmona, 2010).

Generalisation indicates whether the process model is not overfitting to the behaviour in the event log and describes the actual system (Buijs et al., 2014). The generalisation value is calculated through a token-based replay operation, using the formula below where *avg_t* is the average of the inner value over all transitions and *freq(t)* is the frequency of t after the replay (Berti et al., 2023).

$$Generalisation = 1 - avg\_t\left(\sqrt{1/_{freq(t)}}\right)$$

Simplicity quantifies the extent to which a model is easy to interpret and understand by humans and it is calculated as the inverse arc degree, where the average degree of a place/transition is defined as the sum of input arcs and output arcs (Blum, 2015).

$$Simplicity = \frac{1}{1 + \max\left(mean\_degree - k, 0\right)}$$

5.1.2 Value Stream Identification

Feature Generation

Depending on the size of the event log and other characteristics such as routing frequencies, it may be beneficial to reduce the data size and potential noisy behaviours by using sampling and filtering techniques.

Then, using the dataset in Figure 5-3, new features are generated to represent the similarity between individual traces (Zandkarimi et al., 2020). In other words, the new features need to represent the key characteristics of the part and product routings (i.e. traces). For this, trace profiles have been created based on the given event log. Trace profiles are vectors that include a set of items describing a trace from a specific perspective (Song et al., 2009), which in this case is value streams. In this phase, an activity profile is created as a "bag of operations" where every item in the profile vector represents an operation in the event log. This approach is in line with the traditional part-machine matrix used in group technology approaches such as Rank Order Clustering (King, 1980) and more recent methods that consider both activities and transitions (Kong et al., 2018). As the activity profile alone does not represent the sequence of operations that a part goes through, an additional transition profile is created. The transition vector includes the binary combination of operations of a specific part routing. In general, the trace profiles are generated using the formulations below.

Let S be a set of unique product serials, and O be a set of operations.

$S = \{s_1, s_2, .. , s_i, .. , s_n\} \qquad 0 < i < n$

$O = \{o_1, o_2, .. , o_j, .. , o_m\} \qquad 0 < j < m$

Trace profiles are defined by $R = [A \mid T]$, where A is the activity profile and T is the transition profile.

$$A_{ij} = \begin{cases} 1, & if\ product\ serial\ i\ uses\ activity\ j \\ 0, & otherwise \end{cases}$$

$$T_{i\,k-l} = \begin{cases} 1, & if\ product\ serial\ i\ transits\ from\ activity\ k\ to\ activity\ l, with\ k \neq l \\ 0, & otherwise \end{cases}$$

For example, if a part goes through operations $a \rightarrow b \rightarrow c$ the resulting activity and transition profiles are the ones represented below (see Table 5-1).

Table 5-1: Routing profile using activities and transitions

CASE ID	Activity Profile			Transition Profile		
Product ID – Serial no.	a	b	c	a-b	a-c	b-c
P1 - XYZ	1	1	1	1	0	1

Feature Selection

Typically, most products undergo several operations or transformations before being ready to leave the factory. Therefore, the dataset resulting from the feature generation phase is likely to have a significant number of features. Among these, some may be redundant, irrelevant or even misguide the clustering algorithm. To address this issue and improve comprehensibility at the same time, a selection of relevant features needs to be defined (Dy & Brodley, 2004). The features generated through profiling the traces in the event log are filtered using a multi-objective selection approach. Multi-objective optimisation is a common choice for selecting the most representative features (Mierswa & Wurst, 2006). The key objectives are (1) maximising the cluster density and (2) maximising the number of features (Mierswa & Wurst, 2006). Cluster density is measured using the Davies Bouldin index, which is calculated using the formula below. In the formula, n is the number of clusters, c_i is the centroid of cluster i, σ_i is the average distance of all points of cluster i to their centroid, and $d(c_i,c_j)$ is the distance between the centroids of cluster i and j (Davies & Bouldin, 1979).

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq 1}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Maximising the number of features nf_p prevents the clustering algorithm from selecting trivial solutions. The trade-off between Davies Bouldin indices and the number of features is denoted by the pair (DB_p, nf_p) and the different values are represented in Pareto plots (Mierswa & Wurst, 2006). From this representation it is possible to choose the optimal number of features.

To solve the optimisation problem, an evolutionary approach is used. Evolutionary algorithms are population-based approaches that mimic natural evolution (Bartz-Beielstein et al., 2014). Generally, evolutionary algorithms are defined as a "collective term for all variants of (probabilistic) optimization and approximation

algorithms that are inspired by Darwinian evolution. Optimal states are approximated by successive improvements based on the variation-selection-paradigm. Thereby, the variation operators produce genetic diversity, and the selection directs the evolutionary search" (Beyer et al., 2002). Genetic algorithms are a variant of evolutionary algorithms, and they use binary strings inspired by the genetic code in natural life to solve computational problems (Holland, 1973). In feature selection problems, a population is comprised of a set of binary vectors (i.e. indiviauls) representing whether features are selected or not (Hamdani et al., 2007). In particular, the Non-dominated Sorting Genetic Algorithm (NSGA-II) is chosen for this research. The algorithm was developed by Deb et al. (2002) and it has become an established method for multi-objective optimisation problems (X.-S. Yang, 2014). The key principles of NSGA-II include non-dominated sorting, elite-preserving operator, crowding distance calculating, and selection operator (Verma et al., 2021).

According to the principle of non-dominated sorting, the population of individuals is sorted using the concept of Pareto dominance. Given a set of feasible solutions to the multi-objective optimisation problem, when a solution is not dominated by any other one (i.e. a solution is better than any other one), it is called Pareto optimal solution. The objective vector corresponding to the set of all Pareto optimal solutions (i.e. Pareto set) is defined as Pareto front. From the initial population, a first rank is assigned to the non-dominated individuals, which are then removed and placed in the first Pareto front. In the rest of the population, the non-dominated individuals are placed in the second Pareto front with a second rank. The process continues until all individuals are allocated to different fronts with their corresponding ranks. According to the elite-preserving principle, elite solutions are retained and directly moved to the next generation. The crowding distance principle measures the density of solutions surrounding a particular solution and it considers a solution to be in a less crowded region when its crowded distance is large. The population in the next generation is selected using the crowded tournament principle, according to which an individual is chosen over another one of the population if it has better rank or higher crowding distance in case of equal rank (Verma et al., 2021).

The key steps of NSGA-II can be summarised as follows (Verma et al., 2021).

Step 1: A random population P_t of size N is initialised.

Step 2: A new population Q_t is generated from P_t through mutation and crossover operations. Crossover is a form of recombination of two parents, while mutation is a Crossover is performed choosing two individuals from the population with a given probability p_c and combining the corresponding genes (0-1 bits) to form two offspring (Mitchell, 1998). In particular, with uniform crossover, the 1-bits from the parent individuals are uniformly distributed with a chance of 0.5 (Syswerda, 1989). Mutation is the random negation of a bit position of an individual occurring with probability p_m (Beyer et al., 2002).

Step 3: The populations P_t and Q_t are combined to form a new population R_t and the non-dominated sorting is applied to R_t.

Step 4: The population individuals of R_t are ranked into different fronts according to the non-dominated principle.

Step 5: From R_t, N individuals are selected to generate the next population P_{t+1}.

- If the size of the first front is grater or equal to N, then the N individuals are selected from the least crowded region.
- If the size of the first front is less than or equal to N, then all individuals in the first front are moved to the next generation and the remaining individuals in the least crowded region of the second front are added.

Step 6: The populations P_{t+2}, P_{t+3}, P_{t+4}, etc. are generated until the stopping criteria is met.

The tuning parameters chosen for the NSGA-II algorithm used for multi-objective feature selection are summarised in Table 5-2.

Table 5-2: Multi-Objective Feature Selection

Optimisation Objectives	*Objective 1*	Maximise the number of features
	Objective 2	Maximise cluster density
	Population size	Number of features (i.e. number of attributes in the trace profiles dataset)
	Crossover	Uniform crossover with 50% probability

NSGA-II Tuning Parameters	*Mutation*	Mutation with probability [1/(number of features)]
	Stopping criteria	Convergence

Clustering Input

The dataset generated by the application of the previous steps is used to calculate the distance between traces representing routings of parts and products. In fact, the goal of cluster analysis is identifying "natural" clusters by grouping "similar" objects together (Dy & Brodley, 2004). This concept of similarity is expressed as optimisation problem using a specific measure. As the profile vectors contain only binary information, arithmetic distance measures can be used to measure the similarity between routings. Among the various similarity metrics, the Jaccard index is an established and intuitive measurement suitable for comparing binary vectors. Given two sets *S* and *T*, the Jaccard similarity is defined as the intersection divided by the size of the union of the two (Fletcher & Islam, 2018).

$$J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

Clustering

In unsupervised learning tasks, the aim of the cluster analysis is to group data into sets of similar data points. In the context of this research, clusters represent product families, and the data points represent related products. Clustering algorithms can be broadly classified into two main categories, namely hierarchical clustering algorithms, and partitional clustering algorithms (Ezugwu et al., 2022). Depending on the desired level of visibility and user interactivity, different approaches can be used. In this research, three algorithms are evaluated and compared, namely agglomerative clustering, K-Means, and X-Means. One of the key tasks of unsupervised learning is identifying the number of clusters. While X-Means is able to achieve this without any user specification, K-Means and agglomerative clustering require manual input. On the other hand, the first two allow for easier application of domain knowledge KPIs in the selection of the optimal number of clusters.

Table 5-3: Advantages and Disadvantages of clustering approaches for value stream identification

	Advantages	Disadvantages
Agglomerative Clustering	Hierarchically generates value streams and related sub-value streams.	Requires domain knowledge to interpret the data and select partitions. Interpretability may be difficult when numerous products are recorded.
K-Means with fitness function	Enables the selection of optimal partitions based on various domain knowledge KPIs.	Requires domain knowledge to interpret the data and select partitions.
X-Means	Fully automated approach.	Partitions are generated based on product routings only.

The clustering phase results in clusters of product serials (case IDs). Based on the product ID that the serials are associated to, the product families are derived. In some cases, product IDs may have their serials associated to different clusters. Based on the assumption that product routings may have abnormal sequences due to events such as recording errors, re-work operations, or machine failures, the smaller proportion of product serials associated to a different cluster is treated as outlier and removed.

Agglomerative Clustering

Hierarchical, agglomerative clustering is an established unsupervised learning technique (Müllner, 2011). The algorithm takes in input a dataset combined with a dissimilarity index (i.e. 1 – Jaccard similarity) and clusters are formed using a bottom-up approach (Ezugwu et al., 2022; Müllner, 2011). Thus, single objects are iteratively combined into larger clusters based on a similarity or distance metric until all objects are merged into a single cluster. This process results in a dendrogram representing the hierarchical structure of the clusters (Ezugwu et al., 2022). Merging subset of points is determined using a linkage metric that generalises the distance between individual points and subset of points. The three basic linkage metrics are single linkage, average linkage, and complete linkage (Ezugwu et al., 2022). While single linkage (or nearest neighbour) calculates the closest distance from any points of one cluster to any other cluster point, complete linkage measures the longest distance. Average distance

calculates the means of all distances of all data points between clusters (Ezugwu et al., 2022). The formulas for the three basic linkage metrics are presented in Table 5-4, where the distance is calculated assigning for all points i in cluster *u* and *j* in cluster *v*.

Table 5-4: Basic Linkage Metrics

Linkage Metric	**Formula**
Single	$d(u,v) = \min\left(dist(u[i], v[j])\right)$
Average	$d(u,v) = \max\left(dist(u[i], v[j])\right)$
Complete	$d(u,v) = \sum_{ij} \frac{d(u[i], v[j])}{(\lvert u\rvert * \lvert v\rvert)}$

For the identification of key value streams using agglomerative clustering, the chosen distance or dissimilarity measure is 1 – Jaccard similarity, and all three basic linkage metrics are evaluated and compared. From the binary dataset with the trace profiles, a distance matrix for the product IDs is generated by aggregating the corresponding product serials and calculating the mode. Then the matrix is used to generate a dendrogram displaying the hierarchical cluster formation. Based on a pre-defined similarity threshold, it is then possible to identify the key product families. The advantage of this method is that it enables clear visualisation of the main product families as well as sub-families and the associated similarity scores. The disadvantage of this solution is that it requires the user to interpret the results and define a similarity threshold to obtain results. This is particularly critical for companies that manufacture hundreds of products, as the resulting dendrogram would be complex and potentially difficult to interpret.

K-Means

K-Means is an established partitional clustering algorithm that clusters data based on a number of partitions K defined a priori. After the algorithm initialises K centroids, the algorithm iteratively assigns each data point to the nearest centroid and then recalculates the position of the K centroids by taking the mean value of all the data points previously assigned to that centroid until the centroids no longer move (Velmurugan & Santhanam, 2011). The objective function used to choose the centroid is the within-cluster-sum-of-squares.

$$\sum_{i=0}^{n} min_{\mu_j \in j} \left(\left\| x_i - \mu_j \right\|^2 \right)$$

While K-Means has relatively low time complexity and high computing efficiency (D. Xu & Tian, 2015), one of the main shortcomings is that the number of clusters is required to be supplied by the user a priori (Pelleg & Moore, 2000). An established way to overcome this limitation is to generate a fitness function by iterating over different values of K and calculating the corresponding performance metric. A fitness function that considers formal aspects as well as aspects related to a specific application is also called as desirability function (Harrington, 1965). Since the goal of this research is to identify key value streams, domain knowledge KPIs are used to generate a fitness function that enables the selection of the optimal number of partitions. Based on the information contained in an event log, it is possible to calculate the following metrics: routing homogeneity within product families, volume distribution across clusters, product distribution across clusters, operations across clusters, critical operations, and critical product serials. The first three metrics focus on the product families associated to each value stream, while the remaining three focus on the resources used by the product families. Thus, the desirability index W (Harrington, 1965) is defined as:

$$W: \{w_1, w_2, \dots, w_d\} \longrightarrow [0,1]$$

where 0 represent the worst desirability, and 1 the best.

Depending on the use case, that is the purpose for which value streams are being identified (e.g. scheduling problems, layout design, process visualisation), different sets of indices *W* may be selected. Then all indices can be combined using the geometric mean of *W*, as suggested by Harrington (1965).

$$W(w_1, w_2, \dots, w_6) = \sqrt[6]{\prod_{i=1}^{6} w_i}$$

The homogeneity of the product families is calculated as the mean similarity of products within clusters (i.e. average within-cluster similarity). Given *S* the set of product serial numbers, and *K* the set of partitions, the value is calculated as:

$$w_1\left(C^{(k)}\right) = \frac{1}{K}\sum_{k=1}^{K} J_{C_k}$$

Where J_{Ck} is the average pairwise Jaccard similarity of product serials in cluster C_k.

The volume distribution is calculated as the number of product serials per cluster, while the products distribution is calculated as the number of product IDs per cluster. In both cases, the distribution across clusters is calculated using a range-to-total ratio, where 0 indicates perfect evenness. These metrics provide information regarding the size of the product families which is useful for applications such as production levelling and scheduling, for which even-sized product families are preferred (Bohnen et al., 2011).

The volume distribution across clusters is calculated as:

$$w_2\left(C^{(k)}\right) = 1 - \frac{C_{V\,max} - C_{V\,min}}{\sum_{k=1}^{K} V_k}$$

Where $C^{(k)}$is a partition with *k* clusters, $V_k = |S_k|$ is the volume defined as the number of product serials recorded in the partition, and $C_{V\,max} = \arg max_{k\in\{1,..,K\}} V_k$ and $C_{V\,min} = \arg min_{k\in\{1,..,K\}} V_k$ are the maximum and the minimum volumes across all clusters.

Similarly, given *P* the set of product IDs, the product ID distribution across clusters is calculated as:

$$w_3\left(C^{(k)}\right) = 1 - \frac{C_{P\,max} - C_{P\,min}}{\sum_{k=1}^{K} I_k}$$

Where $I_K = |P_K|$ is the number of product IDs recorded in the partition, and $C_{P\,max} = \arg max_{k\in\{1,..,K\}} I_k$ and $C_{P\,min} = \arg min_{k\in\{1,..,K\}} I_k$ are the maximum and the minimum volumes across all clusters. In other words, the volume distribution and the product distribution are calculated as the inverse of the a range-to-total measure across partitions.

The operations across clusters are quantified as a percentage of the total activities that are present in more than one cluster process model. Although this provides an indication of how separate the process models for each cluster are, it does not provide an accurate indication of the value stream independence.

Given *O* the set of operations recorded across partitions, the index is calculated as:

$$w_4\left(C^{(k)}\right) = 1 - \frac{|\{o|o \in O_k \cap O_l\}|}{|O|} \quad \text{where } k, l \in K, \text{and } k \neq l$$

Operations across clusters, is calculated as a the inverse of the percentage of the total operations.

Typically, the application of PFA for the resolution of the cell formation problem has as key objective to minimise inter-cell movements (YounesSinaki et al., 2023), that is minimising the amounts of products using resources allocated to different clusters. Since operations are associated to multiple scanning points, indicating the presence of multiple stations executing the same activity, the presence of the same operation in different clusters does not necessarily imply inter-cluster movements. Therefore, critical operations are defined as the activities that are present in multiple clusters and whose scanning points are less than the number of clusters in which they are recorded. The critical product serials are the ones that use critical activities in a specific cluster configuration, thus causing inter-cluster movements. Similarly, to w_3, the critical operations w_4 and critical product serials w_5 are calculated as a percentage of the total operations and product serials respectively.

Given *O* the set of operations recorded across partitions, *SC* the set of scanning points, and *S* the set of product serials recorded in the event, log, w_4 and w_5 are calculated as:

$$w_5\left(C^{(k)}\right) = 1 - \frac{|\{o|o \in O_k \cap O_l\}|}{|O|} \quad \text{where } k, l \in K, \qquad k \neq l, and\ |SC_o| < |K_O|$$

$$w_6\left(C^{(k)}\right) = 1 - \frac{\left|\{s|s\ uses\ any\ \{K_O\}\}\right|}{|S|}$$

X-Means

X-means builds on the principles of the popular algorithm K-means with some improvements. In particular, X-Means provides a solution to the problem of selecting the optimal number of clusters, and automatically determines the number of clusters. The criteria used to search the space of cluster locations and number of clusters is the Bayesian Information Criterion (BIC) (Pelleg & Moore, 2000). The BIC is a statistical measure for model selection that calculates the trade-off between model fit and complexity (i.e. number of parameters). In the formula below, given a dataset D and a set of alternative models Mj, $\widehat{l_j}(D)$ is the log-likelihood of the data according to the j-th model and taken at the maximum likelihood point, and pj is the number of parameters in Mj (Kass & Wasserman, 1995; Pelleg & Moore, 2000).

$$BIC(M_j) = \widehat{l_j}(D) - \frac{p_j}{2} \cdot logR$$

Overall, X-means enables the generation of clusters based on the Jaccard similarity between profile vectors and returns easily interpretable results. Compared to the previous approaches X-Means has the advantage of fully automating the identification of the optimal number of value streams. The number of clusters is derived using the BIC calculated on the features in the input dataset, which represent the parameters of the trace profiles. Therefore, while this approach offers the advantage of increased speed and full automation, it does not allow for additional considerations regarding the number of value streams (e.g. production volume, critical operations).

5.1.3 Evaluation

The clusters representing the product families are evaluated using statistical performance metrics and Process Mining evaluation metrics. To evaluate the clusters from a Machine Learning perspective, the average within-cluster similarity and the Davies Bouldin index are used. On one hand, the average within-cluster similarity is a quality measure quantifying the homogeneity between product routings within the same cluster. On the other hand, the Davies Bouldin index measures the relationship between intra-cluster and inter-cluster distances. Solutions with Davies Bouldin index values close to 0 are preferred.

The resulting clusters represent the different value streams. Based on the associated products, the original event log can be divided into smaller ones. Using Process Mining techniques, these can be used to map the process models of each value stream, and the process models can be evaluated using replay fitness, precision, generalisation, and simplicity.

Finally, the process models of each value stream can also be evaluated from a domain knowledge perspective by visually inspecting the models and compare them to the physical manufacturing processes.

5.2 EVENT LOG WITH MISSING ATTRIBUTES

5.2.1 Product ID is Missing

The core method presented in chapter 6.1 uses the unique serial number of the different products to create a vector describing their routing. As a result, the clustering algorithm finds partitions representing the product families as distinct groups of product serial numbers. By knowing the association between serial number and product name, it is then possible to derive the clusters of product families identified by groups of product names.

If the product name is not recorded in the event log, the method is still able to identify product families based on the routings of products identified by unique serial number. However, results would not be as informative as the optimal use case in which product families would include a group of product names.

For this case, deriving the relationship between unique serial numbers and associated products is not possible. This is due to the underlaying definition of product families according to the PFA principles. In fact, based on this definition, products belonging to the same product family share highly similar material flows. Therefore, the proposed approach would not be able to distinguish between products within a product family, as by definition the product routings are the same or very similar.

5.2.2 Operation ID is Missing

In case the relationship between scanning points and operations is unknown, the event log needs to be pre-processed before the PFA principles can be applied. In practice, this scenario represents situations in which manufacturing firms are able to collect production data recording products moving around the factory through different scanning points. In this scenario, multiple scanning points may be associated to the same operation, meaning that there are multiple parallel stations dedicated to the same activity.

The event log is used to generate a dataset that contains the product names and the routings representing the serial numbers (i.e. different instances of the products) and their routings. For each product name, the examples in the dataset are filtered down to the most frequent number of operations. For example, in Figure 5-4, product A goes through 3 activities 83% times, therefore only examples of product A containing 3 operations are included.

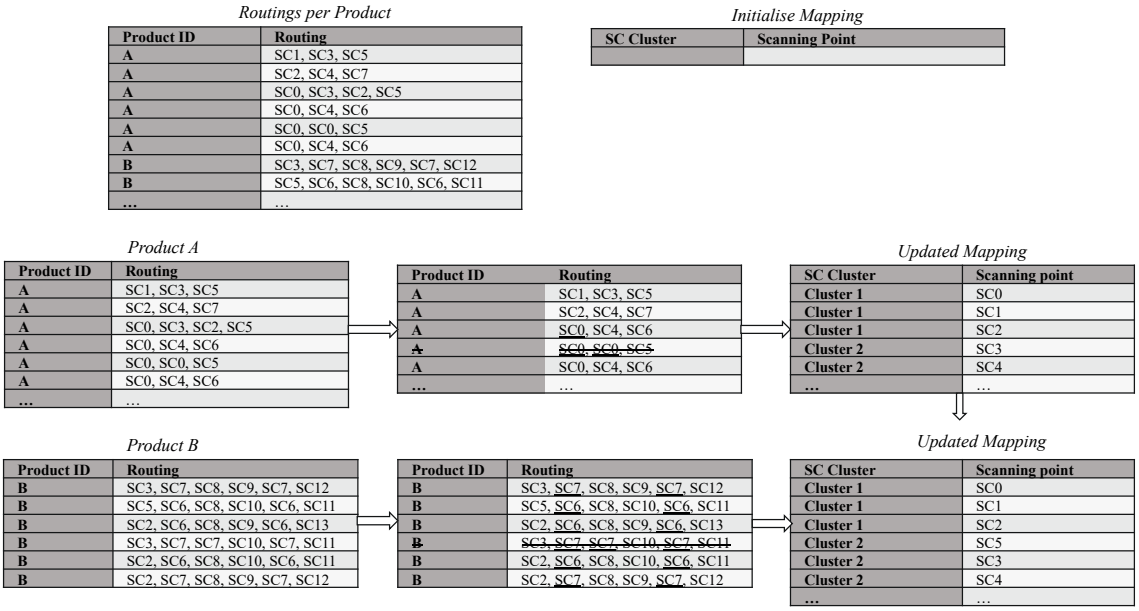


Figure 5-4: Procedure for identifying operations-scanning points associations

Then, for each product name, the scanning points are grouped into clusters representing a specific operation based on the order of occurrence. Thus, for product A, SC0, SC1, SC2 are grouped into a cluster. The key underlaying assumption is that different serial numbers of the same product undergoing a specific number of operations are likely to undergo work operations in the same order. Although this assumption might be valid in many cases, some work operations may occur in a different sequence. For example, after activity 1 a product can undergo activity 2 and then activity 3, or alternatively after activity 1 a product can undergo activity 3 and then activity 2. To reduce the possibility of grouping scanning points incorrectly, the algorithm checks for each product if any of the scanning points is recorded in a different sequence. In case a scanning point is recorded at different positions in the product routings, then the algorithm checks for anomalies. Based on the average occurrence of a specific scanning point, the algorithm attempts to identify and remove anomalies. In the example, for product A, the average occurrence of SC0 is 1, meaning that on average the work operation related to that scanning point is required once by a specific product. However, while most routings record SC0 as first activity, one records it twice. Thus, the algorithm removes the outlier before grouping scanning points based on the order of occurrence. For product B, SC6 and SC7 occur twice on average, meaning that the work operation they are associated to is repeated. Therefore, only the routing in which the scanning point is recorded more than twice is removed before clustering. The algorithm starts mapping scanning points starting from product

with the shortest routing (i.e. products undergoing the smallest number of operations) and identifies clusters of scanning points progressively product-by-product based on the steps explained above. Each cluster is mapped in a new dataset that identifies this relationship. When the algorithm analyses a different product and the related routings, it checks if any of the scanning points has already been mapped. In this case, the cluster including the scanning point already mapped gets associated to the cluster in the mapping. In the example, SC5 will be associated to cluster 2, since product B undergoes SC2, SC3, and SC5 for the first work operation and SC2 and SC3 were previously assigned to the same cluster.

The algorithm continues to analyse product routings progressively and updates the dataset accordingly. In case a product only records two routings, or the sequence lengths have equal frequencies (e.g. for a given product 50% of sequences record two activities, 50% sequences record three activities), the algorithm ignores that product and moves to the following.

5.3 EVENT LOG CLEANING AND FILTERING

The ability to collect quality production data depends on a company's infrastructure, and ultimately on their level of digital maturity. Event logs that are not excellent quality may have issues related to event attributes. These include the following (Bose et al., 2013; van der Aalst, 2016).

- Missing attributes: the entity has occurred in reality but has not been recorded in the event log, therefore an attribute has not been recorded for a specific event.
- Incorrect attributes: the recorded value for a specific attribute is wrong, for example it may refer to a different case.
- Imprecise attribute: the value of a specific attribute is not informative enough, for example it may be too coarse-grained.

Typically, in Manufacturing, event logs may record a variety of attributes. Based on the requirements of Process Mining techniques as well as the principles of PFA, the most important entities are case, activity, and timestamp. In the context of a production event log, these are serial number, operation ID, and timestamps, respectively.

PRODUCT_NAME	SERIAL_NO	SCANNING_POINT	OPERATION_ID	TIMESTAMP
P1	XYZ	SC2	OP_1	29/11/23 10:05
P1	XYZ	SC4	OP_2	29/11/23 10:23
P1	XYZ	SC1	OP_3	29/11/23 11:21
..	..	..	..	..
P5	XXX	SC4	OP_2	30/11/23 16:11
P5	XXX	SC3	OP_3	30/11/23 17:22

Figure 5-5: Example of an event log in Manufacturing

Table 5-5 provides a detailed overview of the quality problems for the most important entities in an event log recording production processes.

Table 5-5 - Possible quality problems in an event log (Bose et al., 2013, van der Aalst, 2016)

	Missing attribute	Incorrect attribute	Imprecise attribute
Case: serial number	The event does not refer to a case.	The event refers to the wrong case.	The event may be related to multiple cases due to the ambiguity.

Activity name: operation ID	The event does not refer to an activity.	The event refers to the wrong activity.	The event is too coarse-grained (e.g. multiple activities recorded as one).
Timestamp	The event has no timestamp.	The event as an incorrect timestamp.	The event is too coarse-grained (e.g. day only)
Any attribute: scanning point or product name	The event has a missing attribute.	The event has the incorrect value for the attribute.	The event may be related to multiple attributes.

As mentioned in section 4.2, in practice, the presence of some quality issues may not always be obvious, especially for infrequent behaviours. Based on this observation, different policies are established (see Table 5-6).

The presence of missing attributes can be verified directly due to the absence of specific values. Thus, if missing attributes are detected, the affected traces can be removed (Pegoraro & van der Aalst, 2019). In fact, Process Mining algorithms generate a process model describing the behaviour of various entities (i.e. products) based on a sample. Each product ID has numerous unique serial numbers associated. In other words, for each product type, there are multiple unique instances being manufactured on a regular basis. If some of these routings are filtered out of the event log, the clustering method would use different instances of the same product ID to define product families. Therefore, the behaviour should not change when a different sample of the same process is used (Bose et al., 2013), and the process model of the whole factory, as well as the process models of the different value streams should not change significantly.

As opposed to missing attributes, detecting the presence of incorrect values is more challenging, especially when the occurrence is infrequent. Incorrect data is typically caused by recording errors during the data collection such as faults of the information system or human error during data entry (Pegoraro & van der Aalst, 2019). Logging errors may cause situations in which instances of a process are incorrectly recorded as instances of a different one, incorrect operations are recorded in the traces, or the timestamps do not match the exact times at which products have undergone

certain activities. One of the potential risks is that causal relationships may be represented incorrectly (i.e. “A” then “B” vs. “B” then “A”) (Bose et al., 2013). More generally, the impact of incorrect values is twofold. On one hand, the discovery algorithm may be misled and generate incorrect process models. On the other hand, the identification of product families may be less accurate. Most existing research handles incorrect behaviours by analysing path frequencies, and available solutions include filtering or repairing incorrect values (Conforti et al., 2017; Pegoraro & van der Aalst, 2019; Sani et al., 2018; Wang et al., 2015). Since the presence of incorrect values is not obvious, filtering out infrequent behaviour at the start can prevent inaccurate results and reduce the complexity of process models. Thus, for each product, the routings with the lowest frequency are treated as outliers and filtered out before applying the Process Mining-based method for PFA. Similar to the solution for finding associations between scanning points and operations, the assumption is that due to technological constraints, most products require a fairly consistent sequence of operations, and exceptions may be due to infrequent events such as re-work operations or equipment failures.

Issues related to imprecise data often relate to its coarseness. Information systems may have limited data collection and recording capabilities and therefore they may be able to record limited information, such as the date but not the time or only certain operations. This is also typical of processes that are recorded manually and then digitalised (Pegoraro & van der Aalst, 2019). In contrast to missing and incorrect data, the uncertainty generated by coarse data can be addressed only by improving the data collection systems and procedures, for example by retrofitting existing equipment. For this reason, the impact of imprecise attributes is only evaluated for the scope of this research.

Table 5-6: Policies for managing quality issues

	Policy	Applicable
Missing Attributes	Remove affected traces	When missing values are detected
Incorrect Attributes	Filter out infrequent routings	Always
Imprecise Attributes	Evaluate the impact of coarse information	When using information systems with limited data collection capabilities

Chapter 6: Method Evaluation

Sections of this chapter have been published in the following conference proceedings:

Tomidei, L., Sick, N. & Mathieson, L. (2024). Data-Driven Value Stream Analysis Using Process Mining And Machine Learning. In 51st International Conference on Computers and Industrial Engineering (CIE51). Sydney, Australia.

Tomidei, L., Sick, N., Deuse, J. & Guertler, M. (2023). Extracting Key Value Streams using Process Mining and Machine Learning. In IEEE Conference on Engineering Informatics. 2023 IEEE Engineering Informatics, 1–7. https://doi.org/10.1109/IEEECONF58110.2023.10520644

The evaluation of the methodology presented in chapter 6 aims to cover all possible use cases, as defined in chapter 4.2. The event log used is complete, and although the presence of incorrect values is unknown, there are no missing values, and the data has appropriate granularity. First, the core methodology is presented using the entire dataset. Second, the pre-processing algorithm is evaluated by simulating a scenario in which only scanning points are recorded in the event log. Third, the performance of the methodology is evaluated in case of missing, incorrect, and imprecise attributes by artificially altering 10% of the events.

The implementation has been done using the Process Mining for Python (PM4Py) library, which provides algorithm customisation and integration with other state-of-the-art data science libraries (Berti et al., 2019), as well as RapidMiner.

Case Study Background

The event log used for the evaluation has been provided by a European manufacturer that produces electronic components. This event log includes 57 different products and over a million unique components and it records the movement of these components as they pass through various work operations, with a total of 34 distinct work operations recorded. The majority of these products are produced in low volumes, with less than 5000 pieces per year (see Figure 6-1).

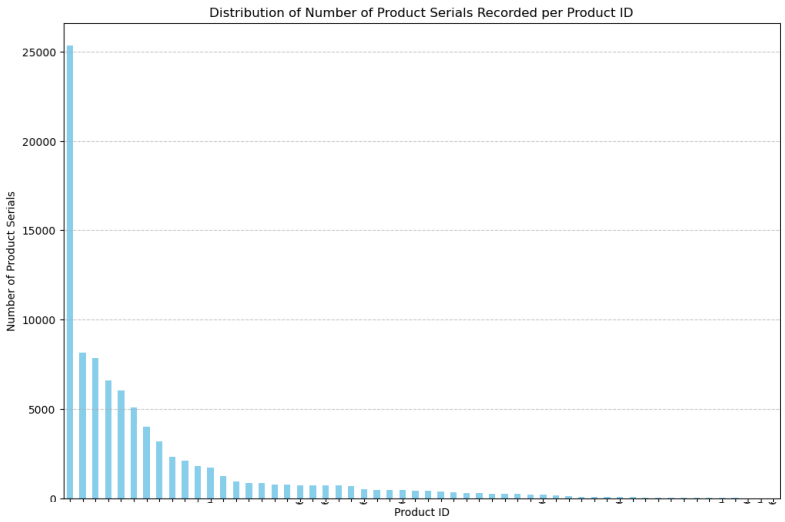


Figure 6-1: Volume distribution over a year

6.1 OPTIMAL EVENT LOG

The event log includes several attributes, and it provides data related to a one-year period. In addition to the ones represented in Figure 6-3, information about the order number, the specific production line, the batch number, and batch status is also included. Firstly, from the original event log, that includes over 1 million events, the cases associated with only one activity have been removed. These represent the parts that have gone through a single processing operation in the factory.

As it is unknown whether the data contains incorrect values, the event log has been further reduced in size by filtering out routings with low frequency. For each product, only the routings corresponding to the top 20% percentile are kept in the event log, while the others are treated as outliers (see examples in Figure 6-2). This approach is in line with the policies for handling data quality issues defined in Table 5-6.

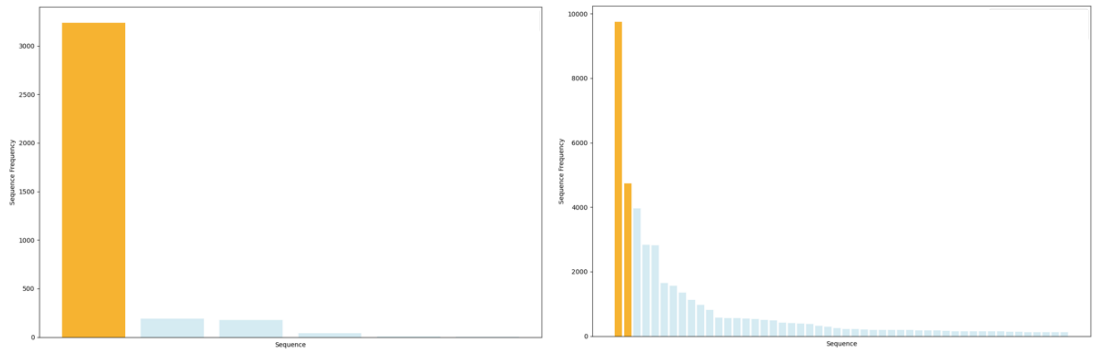


Figure 6-2: Example of routing frequency distribution for 2 products in the event log

6.1.1 Process Model Generation and Evaluation

First, the event log is fed into a Process Mining discovery algorithm to generate a process model of the entire factory. The chosen algorithm is inductive miner - infrequent, filtering infrequent events is set to 30%. The resulting process model using the BPMN representation is shown in Figure 6-3.

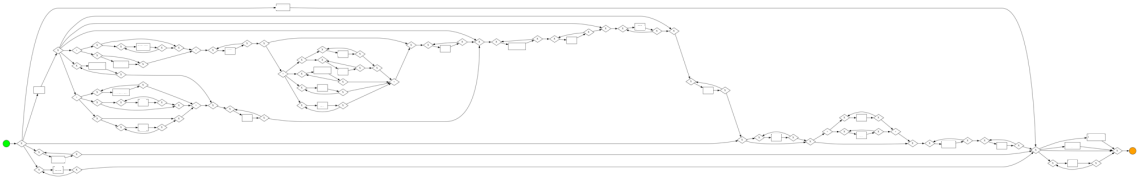


Figure 6-3 - Process Model of the factory floor

The process model is evaluated using the key four quality measures (i.e. replay fitness, precision, generalisation, and simplicity). While most of the evaluation metrics return good results, the simplicity score supports the apparent complexity of the process model.

Table 6-1: Process Model Quality

Metric	Value
Replay fitness	94.4%
Precision	67.7%
Generalisation	88.9%
Simplicity	65.2%

In addition to the process model generated by the inductive miner, the event log is also processed through a discovery algorithm that generates a Directly-Follows Graph (DFG), which is a model where the nodes represent activities, and the edges

directly follows relationships, and is able to represent frequencies (see Figure 6-4) and durations (see Figure 6-5) (van der Aalst, 2019a).

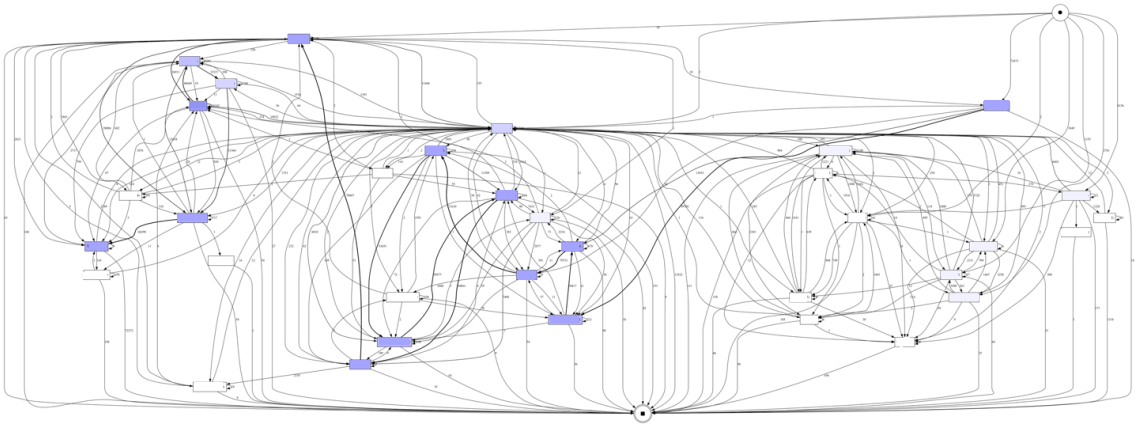


Figure 6-4: Directly-Follows-Graph of the entire factory with path frequencies

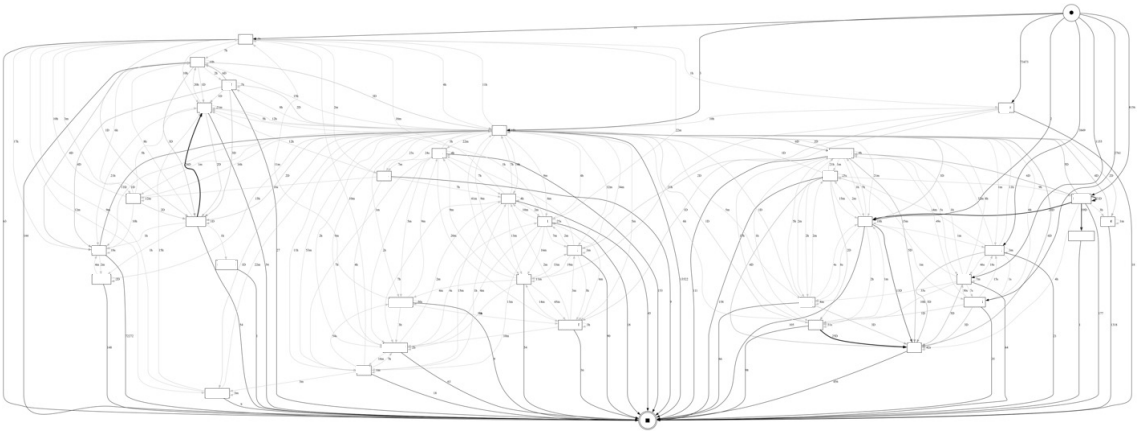


Figure 6-5: Directly-Follows-Graph of the entire factory with path duration

6.1.2 Value Stream Identification

Feature Generation

Filtering out infrequent traces allows to reduce the event log size by two thirds, from 1 million events to 360,073 events. The remaining dataset is sampled using a stratified approach. When handling big or complex event logs, strategies such as sampling can be helpful (Leemans et al., 2015). To do this, two new attributes are created, one representing the trace (i.e. routing) of each part (e.g. a→b→c), and the other one representing the frequency of each specific routing. Then, a 40% stratified sample is taken from each trace. By doing this the size of the event log is reduced while maintaining the distribution of the various traces and reducing noise. Thus, the event log size is reduced to 143,838 events, capturing the movements 57 products and over 16,000 unique serial numbers.

This event log is processed to generate the activity and transition profiles. As a result, a dataset with 87 attributes is produced. Each example represents a trace identified by a specific product identified by a unique serial number.

Feature Selection

The 87 attributes generated from the previous phase result in a dataset characterised by high dimensionality. To reduce complexity and improve interpretability, the features generated through profiling the traces in the event log are filtered using a multi-objective selection approach. The optimisation problem for selecting features is based on the two non-conflicting objectives of (1) maximising cluster density measured by the Davies Bouldin index and (2) maximising the number of attributes. A genetic algorithm has been used to solve the multi-objective optimisation problem and the results are shown in the Pareto front in Figure 6-6.

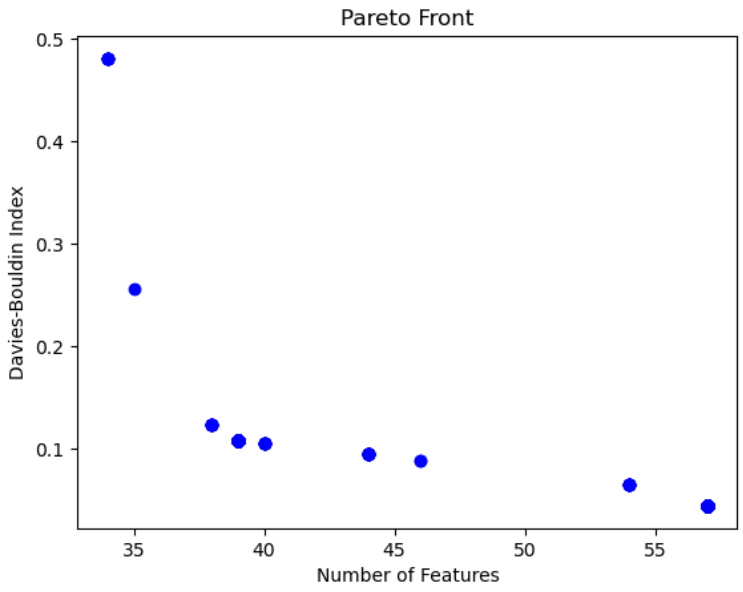


Figure 6-6: Pareto front for feature selection using an optimal event log

Each point in the chart represents a feature set associated with a specific Davies Bouldin Index. The optimal feature set can be determined by the point in the Pareto part with the Davies Bouldin index closest to 0. Thus, the multi-optimisation approach for feature selection reduces the number of features from 87 to 57.

Clustering Input

The new dataset is comprised of 14,617 traces corresponding to product serials and 57 features. Twenty of these features represent individual work operations, while the remaining represent transitions between operations. In the cluster analysis, profile

vectors are compared using the Jaccard distance. From the similarity matrix in Figure 6-7, it is possible to visually notice two main product families.

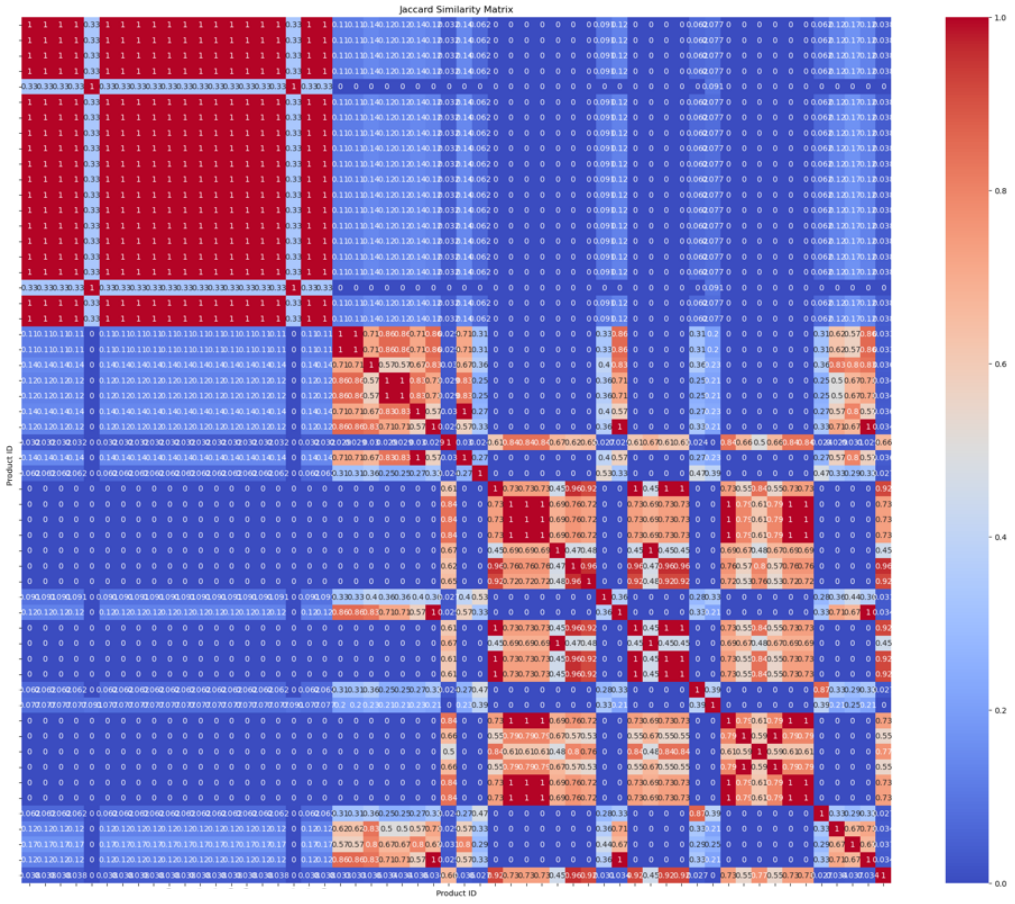


Figure 6-7: Similarity matrix using Jaccard

Clustering

Agglomerative Clustering

The dendrogram generated from the agglomerative clustering algorithm returns a structured hierarchy of product families. Based on the definition of similarity threshold, different clusters can be identified. Assuming a threshold of 20% similarity, three to four product families are identified, depending on the linkage approach (see Figure 6-8).

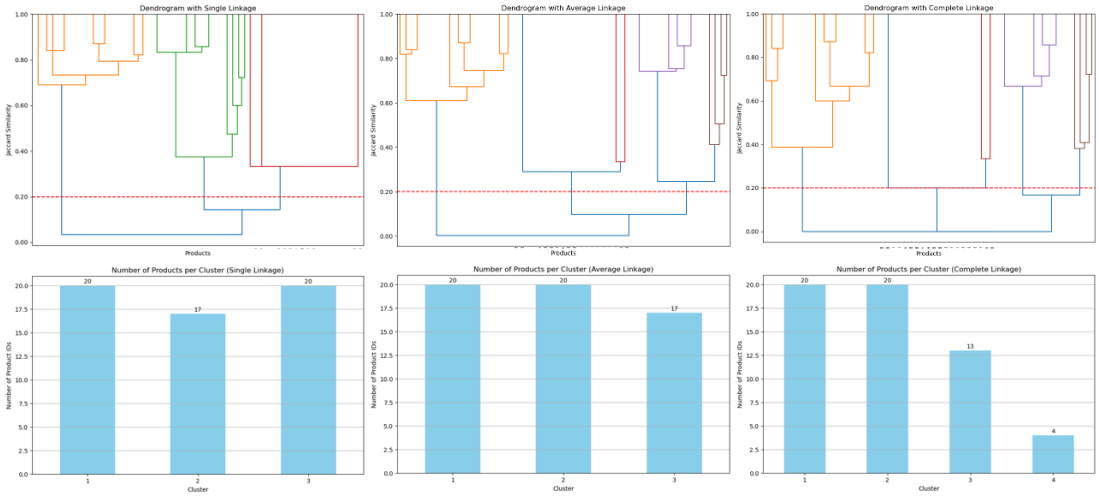


Figure 6-8: Dendrograms generated by agglomerative clustering (left: single linkage, centre: average linkage, right: complete linkage)

K-Means and X-Means Clustering

When applying K-Means clustering the number of partitions k needs to be defined by the user. As mentioned in section 5.1.2, by using a set of desirability functions, it is possible to identify the optimal value of partitions k. The set of desirability functions to be considered ultimately depends on the application for which value streams need to be identified. For the evaluation of this case study, all functions have been included, assuming the identification of value streams has general purpose.

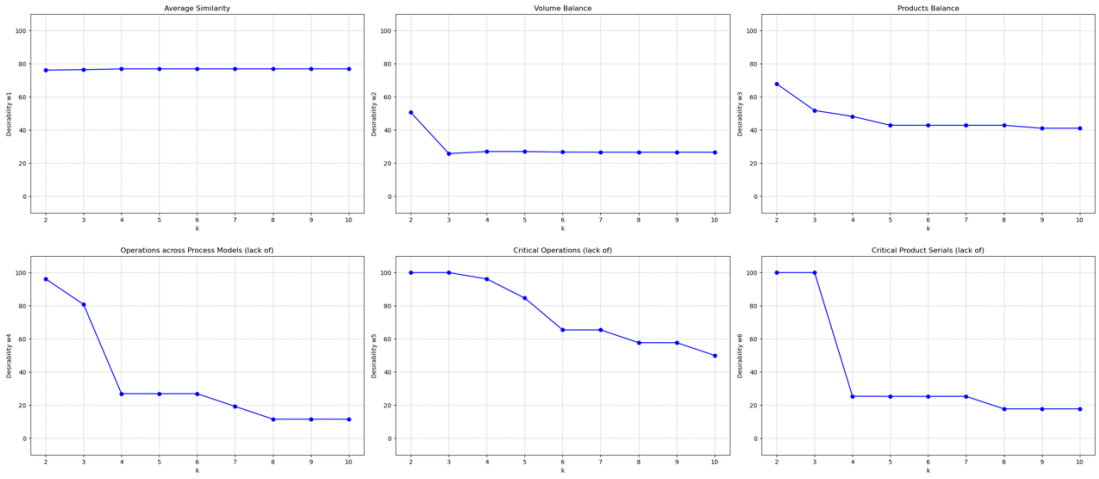


Figure 6-9: Desirability functions w_1, w_2, w_3, w_4, w_5, w_6

The average similarity within clusters remains steady across the possible partitions, with values close to 80%. Both the volume imbalance and the product family imbalance increase for partitions above 2. The number of operations present in

multiple process models also increases from partitions above 2, and more significantly for partitions above 3. The number of critical operations, that is the number of operations that can cause inter-cellular movements, increases only marginally for partitions above 3 and more significantly for partitions above 4. However, the marginal increase in critical operations triggers the inter-cellular movement of a significant number of products for partitions above 3.

The similarity functions also explain the solutions provided by the agglomerative clustering algorithms, for which clusters are generated hierarchically based on intra-cluster similarity. Accordingly, the average similarity w1 slightly improves for partitions above two, but the consideration of additional elements (w2 to w6) reveals that selecting two partitions is more appropriate.

The six desirability functions shown in Figure 6-9 are combined to calculate the desirability index W which directly informs the decision of the best number of partitions (see Figure 6-10).

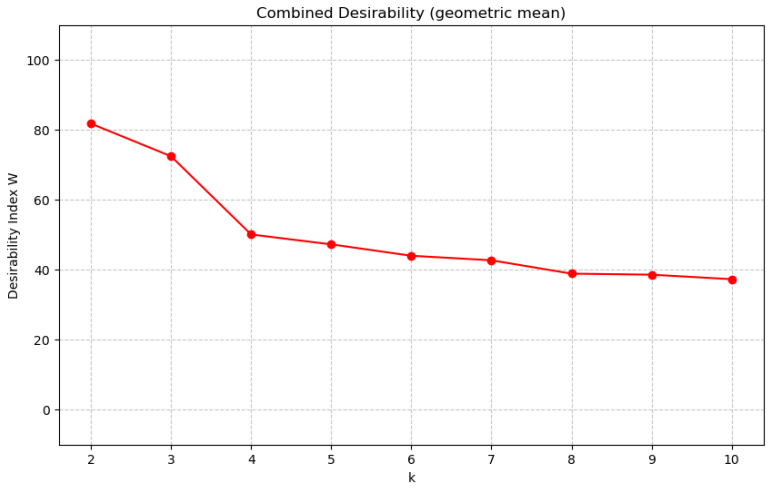


Figure 6-10: Desirability index W calculated as mean of the functions w_1, w_2, w_3, w_4, w_5, w_6

By selecting 2 partitions, the K-Means algorithm returns two product families with 37 and 19 products each.

The same partitions are generated by the X-Means algorithm, which measures the distance between the routings represented by the profile vectors and determines the optimal number of product families as well as the which products belong to each family.

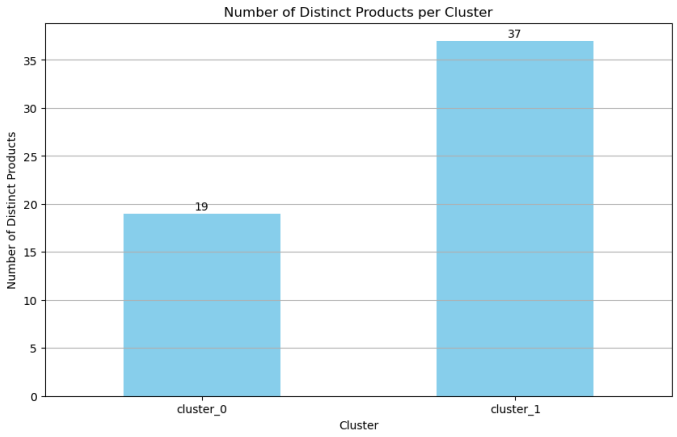


Figure 6-11: Product families generated by K-Means and X-Means

6.1.3 Evaluation

The quality of the clusters generated by the agglomerative clustering approaches is influenced by the similarity threshold applied to the dendrogram. In the use case, the threshold was set to 20% to include product families in which products have routing similarities equal or above that value. The clusters returned by the algorithm display average similarities well above 20%, with values ranging from 49.4% to 80.2% depending on the value stream and the linkage method, thus indicating a satisfactory solution in terms of product family formation. However, the process models generated by the product family configurations score low quality, with some precision values as low as 14% (see Table 6-3).

Compared to the solutions generated by agglomerative clustering, the ones returned by K-Means (K=2) and X-Means provide better results both in terms of clusters and process models. The average similarity within clusters indicates high routing homogeneity within the product families (see Table 6-2).

Table 6-2: Evaluation of the solution generated by K-Means (K=2) and X-Means

		Value Stream 1	Value Stream 2
Clusters Evaluation	Davies Bouldin Index	0.567	
	Average Similarity	96.4%	99.4%
Process Models Evaluation	Replay fitness	97.8%	91.9%
	Precision	86.5%	72.9%
	Generalisation	93.4%	66.2%
	Simplicity	86%	65.7%

Overall, the evaluation of the process models indicates that the process models have less complexity compared to the one representing processes in the entire factory (i.e. before value stream identification). While the process model of value stream 1 is good quality, with high values for fitness, precision, and generalisation, the process model of value stream 2 has lower generalisation and simplicity. As generalisation represents the ability of the model to represent future behaviour, results for value stream 2 indicate that although the model is a fairly accurate representation of the present material flows, the presence of multiple paths may prevent future cases to fit in the existing model (Buijs et al., 2012). The process models discovered by the heuristic miner discovery algorithm, which is considered a generalising algorithm (Buijs et al., 2012), can be found in Appendix C.

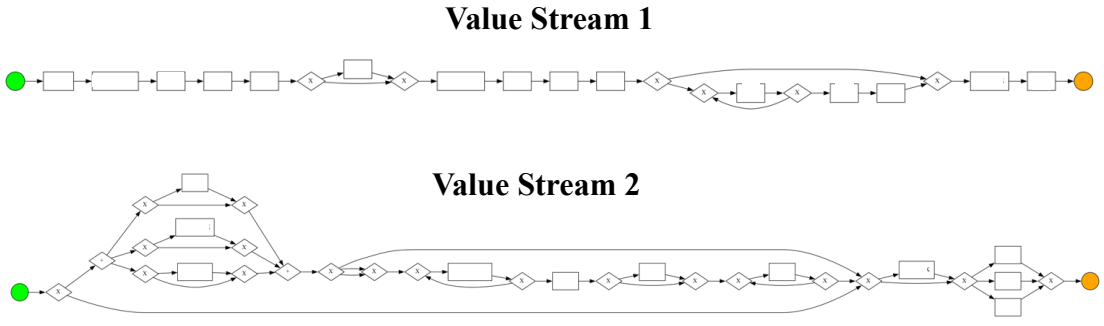


Figure 6-12 - Process models (BPMN) generated by K-Means and X-Means[1]

The centroid table plots the values for the cluster centroids for each attribute, with attributes corresponding to individual work operations or sequence between operations (see Figure 6-13). From this, it is possible to verify which work operations

are unique to each value stream, ultimately enabling better visibility. From a domain knowledge perspective, this is particularly important as it enables direct interpretation of results and the evaluation of the solution.

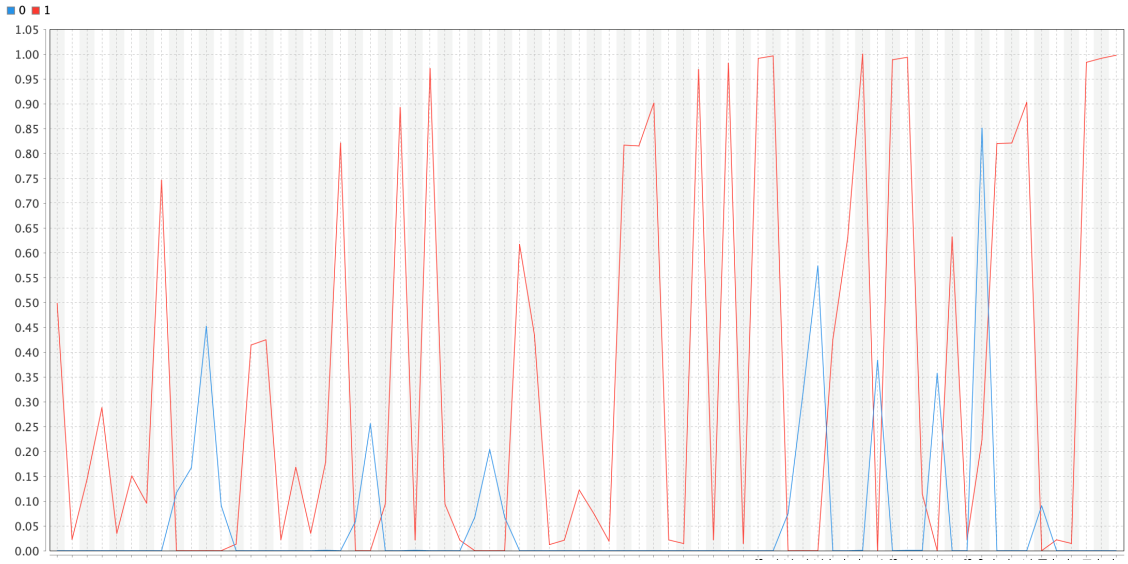
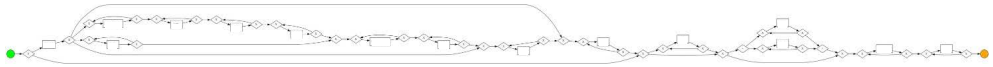
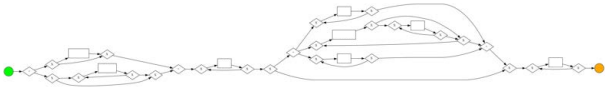
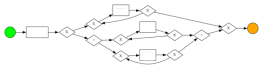
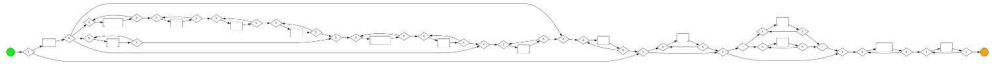
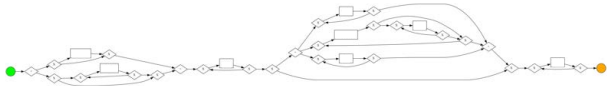
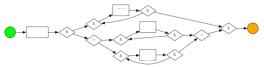
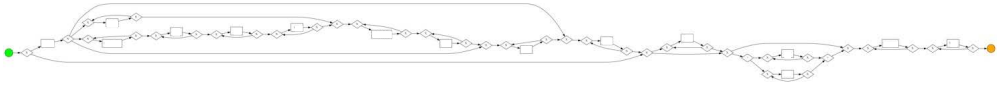
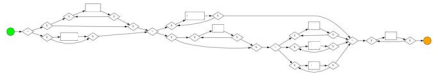
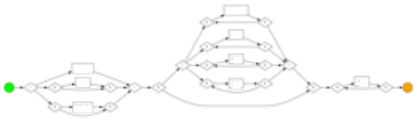


Figure 6-13: Centroid chart[1]

[1] The attributes names indicating operations names and codes in the x-axis have been removed to maintain confidentiality. The same applies for the activities in the process models.

Table 6-3: Process Models (BPMN) generated by agglomerative clustering and evaluation

		Process Model	Average similarity within cluster	Replay Fitness	Precision	Generalisation	Simplicity
Single Linkage	*Value Stream 1*		72.2%	95.1%	54.6%	50.6%	64.7%
	Value Stream 2		58.9%	84.7%	46.7%	24%	64.7%
	Value Stream 3		80.2%	81.2%	14.4%	10.8%	64.7%
Average Linkage	*Value Stream 1*		72.2%	95.1%	54.6%	50.6%	64.7%

	Value Stream 2		58.9%	84.7%	46.7%	24%	64.7%
	Value Stream 3		80.2%	81.2%	14.4%	10.8%	64.7%
Complete Linkage	*Value Stream 1*		72.2%	95.1%	54.6%	50.6%	64.7%
	Value Stream 2		80.2%	81.2%	14.4%	10.8%	64.7%
	Value Stream 3		82.7%	85.1%	42.3%	23.4%	64.7%
	Value Stream 4		49.4%	81.2%	33.9%	21.8%	64.7%

6.2 EVENT LOG WITH MISSING ATTRIBUTES

6.2.1 Case Attribute Missing: Product ID

The core method presented in section 6.1 uses the unique serial number of the different products to create a vector describing their routing. As a result, the clustering algorithm finds partitions representing the product families as distinct groups of product serial numbers. By knowing the association between serial number and product name, it is then possible to derive the clusters of product families identified by groups of product names.

If the product name is not recorded in the event log, the method is still able to identify product families based on the routings of products identified by unique serial number. However, results would not be as informative as the optimal use case in which product families would include a group of product names.

For this case, deriving the relationship between unique serial numbers and associated products is not possible. This is due to the underlaying definition of product families according to the PFA principles. In fact, based on this definition, products belonging to the same product family share the highly similar material flows. Therefore, the proposed approach would not be able to distinguish between products within a product family, as by definition the product routings are the same or very similar.

6.2.2 Activity Name Missing: Operation ID

For complete event logs, the process model can be discovered by using activity names as classifier, with traces corresponding to a sequence of activities, thus enabling the definition of operation sequences for each product in the factory. However, in this scenario, activities (i.e. work operations) are not recorded in the event log. Therefore, to generate a process model of the factory, the scanning points can be used as event classifier.

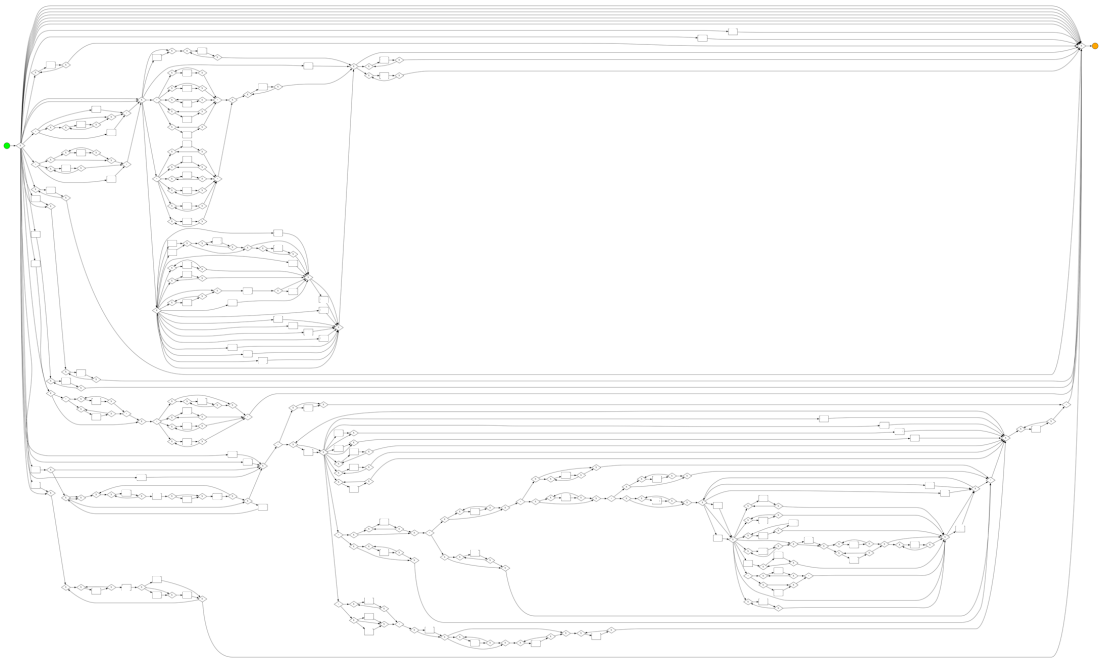


Figure 6-14: Process Model (BPMN) of the factory using scanning points as classifier

The process model represented as a BPMN is discovered using an inductive miner infrequent. The model appears more complex than the one generated using work operations as classifier, as the simplicity score indicates (see Table 6-4). This is due to the fact that most work operations have multiple scanning points associated. Additionally, while most scanning points are uniquely associated to specific operations, a few are associated to multiple operations (see Figure 6-15).

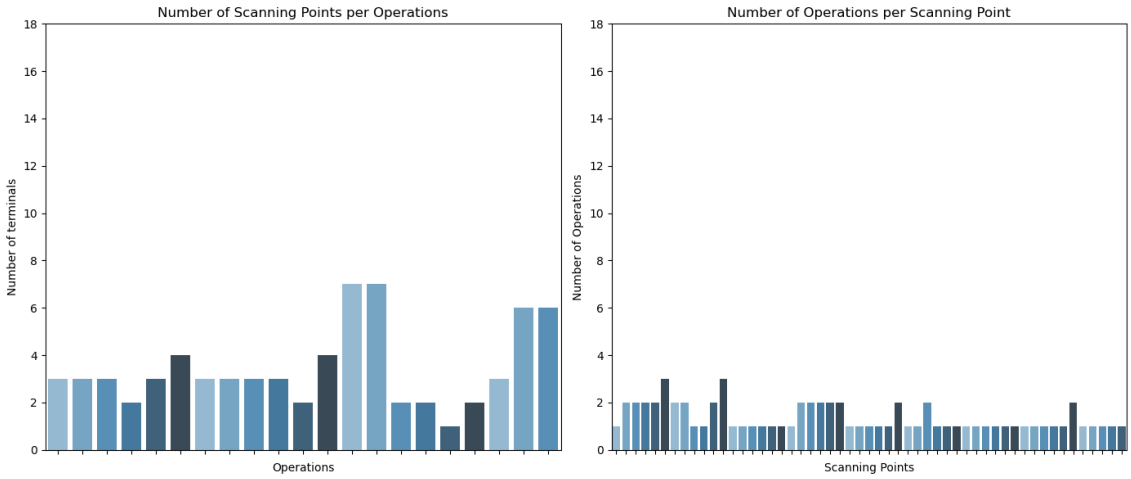


Figure 6-15: Associations between scanning points and operations

In addition to the low simplicity, the precision score indicates that the process model is underfitting the event log, thus representing behaviours that are not present in the data.

Table 6-4: Process Model Evaluation Metrics

Metric	Value
Replay fitness	97.4%
Precision	20.9%
Generalisation	85.6%
Simplicity	58.8%

Event Log Pre-Processing

Firstly, the event log is reduced by removing product serials undergoing one operation and reducing the observation period to one year. The resulting event log has 34 distinct work operations. By applying the algorithm described in the previous section, it is possible to obtain the results shown in Figure 6-16. The event log pre-processing results in 30 clusters and this is caused by two factors. On one hand, the work operations-scanning points associations are not exclusive, as some scanning points are associated to multiple work operations (see Figure 6-15). On the other hand, the serial numbers associated to some product IDs exclusively use certain scanning points. This means that the algorithm is not able to cluster scanning points based on the work operation sequence for those products, and when the algorithm progressively checks new products, it considers the clusters as separate.

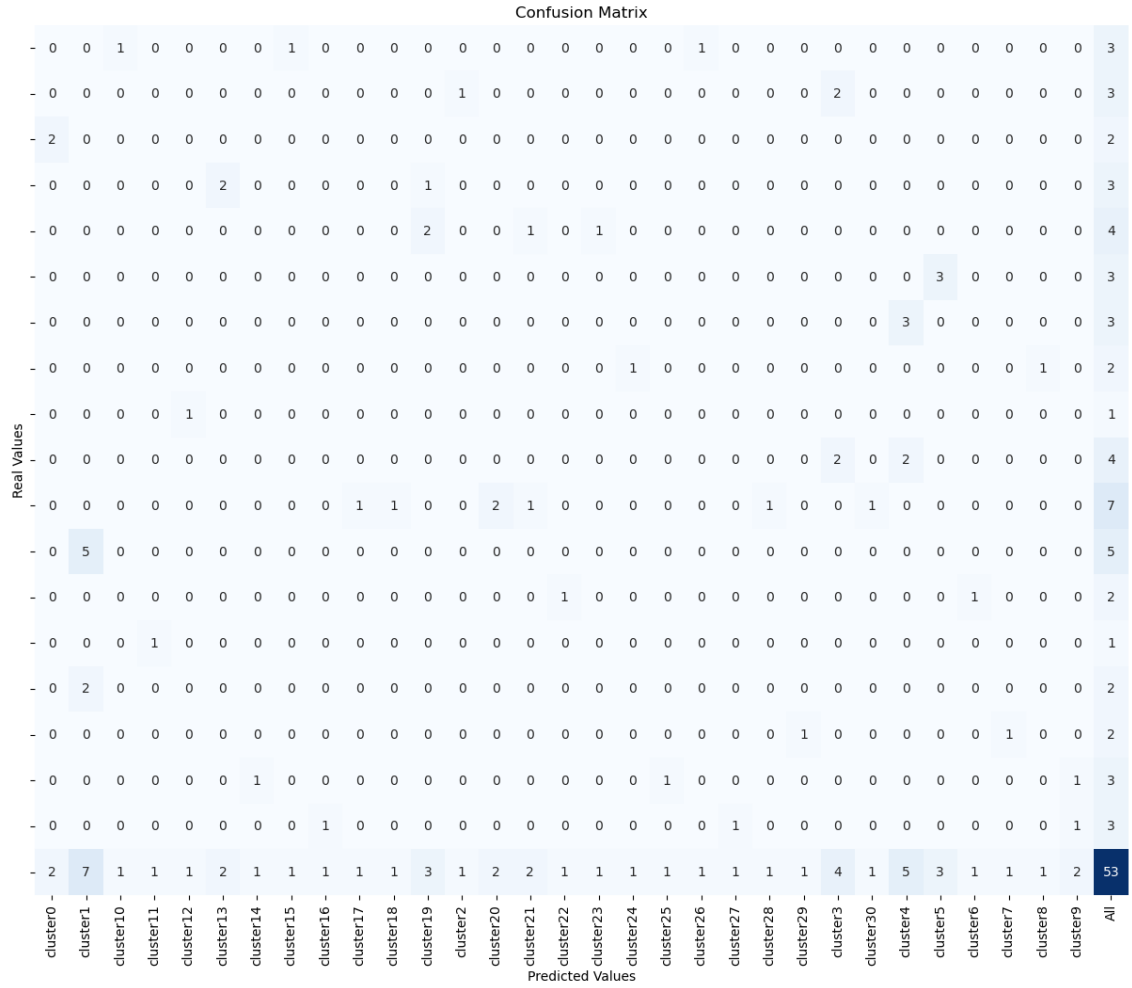


Figure 6-16: Scanning Point Clustering

Based on these associations, it is possible to re-discover the process model of the entire factory.

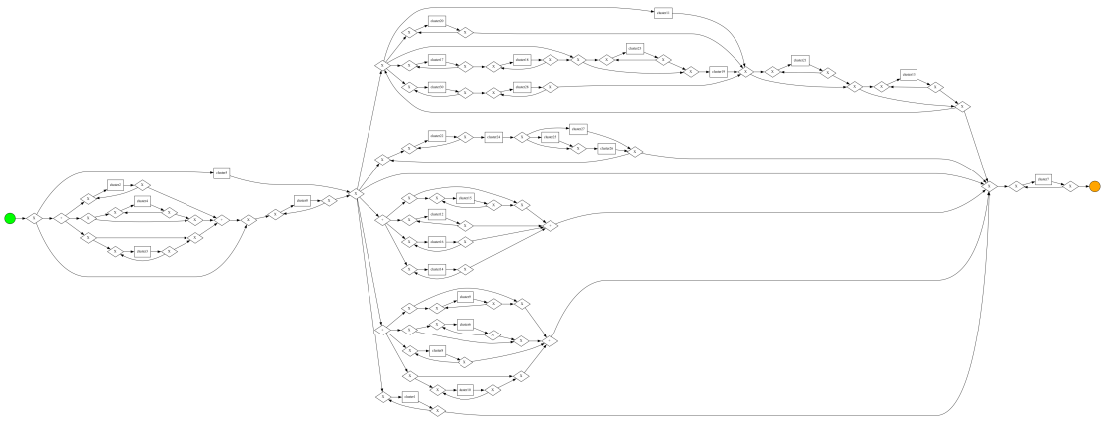


Figure 6-17: Process Model (BPMN) of the factory after pre-processing the event log

Compared to the process model discovered using scanning points as classifier, the process model in Figure 6-17 appears less complex and easier to interpret. This is reflected by the evaluation metrics (see Table 6-5), that indicate a substantial improvement in regard to all quality aspects, particularly precision.

Table 6-5: Process Model Evaluation Metrics

	Scanning Points	Clustered Scanning Points
Replay fitness	97.4%	90.7%
Precision	20.9%	73.4%
Generalisation	85.6%	92.1%
Simplicity	58.8%	63%

Value Stream Identification

In this scenario, since the operation attribute is not recorded, the attributes created to enable the sampling process are two, one for the trace of each part represented by the sequence of scanning point clusters (e.g. cluster_terminal_1→ cluster_terminal_2→ cluster_terminal_3), and the other one representing the frequency of each routing. Because the clusters of scanning points do not fully match with the work operations, the number of different traces per product is higher, thus requiring a re-evaluation of the threshold used for filtering out infrequent traces (see Figure 6-18). Instead of taking the traces from the top 20% of frequency values, as for the use case with operations as classifier, the traces corresponding to the top 10% frequency are retained. This accounts for the additional variability created from the clustering of scanning points, while retaining only the most frequent routings for each product.

The subsequent steps are the same used for identifying key value streams from an optimal event log. The trace profiles for product serials are generated from the filtered event log, creating a binary dataset with several attributes representing activities (i.e. clustered scanning points) and transitions between activities. The optimal subset of features is selected using a genetic algorithm that executes a multi-objective approach aiming to maximise cluster density and maximise the number of features.

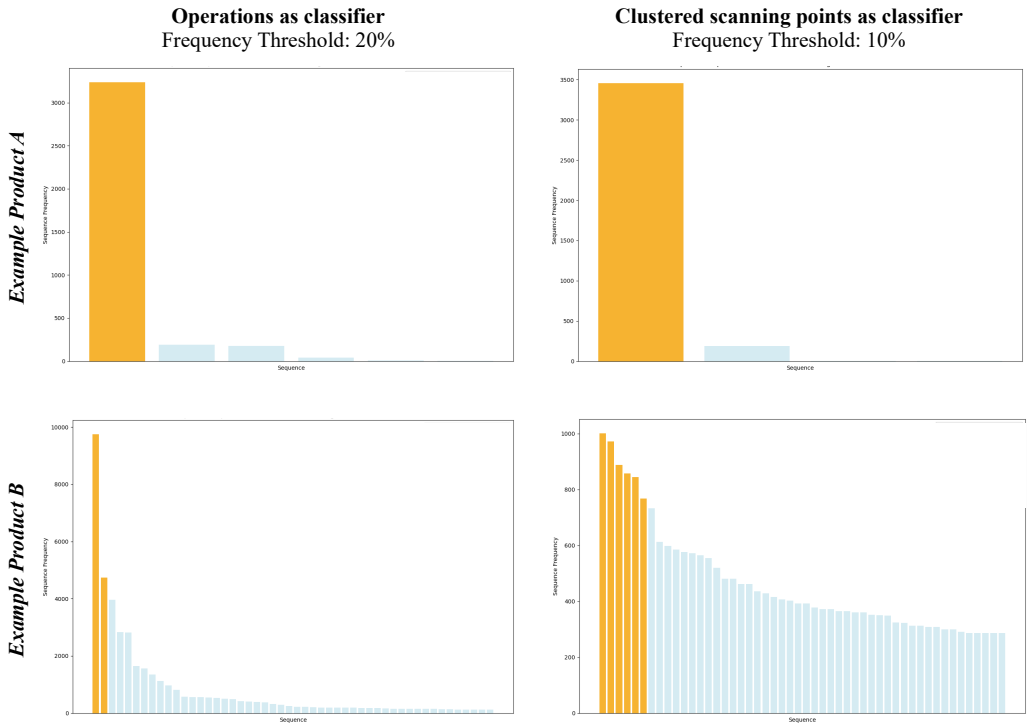


Figure 6-18: Event log filtering after clustering scanning points

By applying the X-means algorithm it is possible to identify the clusters representing the product families. The algorithm returns a product family formation that matches the one generated by the optimal event log, including 2 product families with 37 and 19 products each.

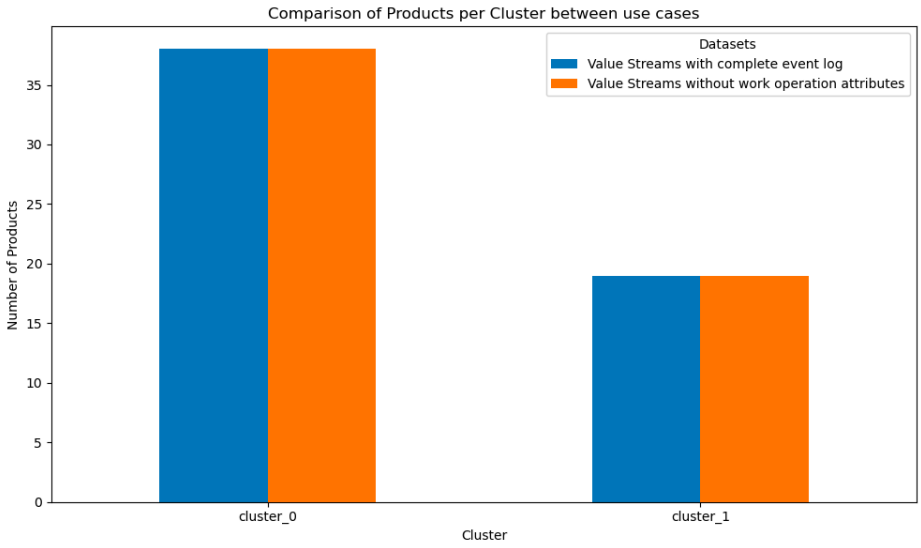


Figure 6-19: Comparison between product families identified using an optimal event log and using an event log without work operation attributes

Evaluation

By clustering scanning points into operations using the algorithm defined in chapter 5.2.2, it is possible to identify key value streams with a clustering performance close to the one generated by using a complete event log. While the Davies Bouldin Index shows minor deviations from the optimal use case, the similarity within products for value stream 2 is 14% lower (see Table 6-6).

Table 6-6: Evaluation comparison between value streams generated using complete event log and a pre-processed event log

		Optimal Event Log		*Pre-Processed Event Log*	
		Value Stream 1	**Value Stream 2**	**Value Stream 1**	**Value Stream 2**
Clusters Evaluation	Davies Bouldin Index	0.567		0.549	
	Average Similarity	96.4%	99.4%	98.9%	84.9%
Process Models Evaluation	Replay fitness	97.8%	91.9%	99.1%	98.7%
	Precision	86.5%	72.9%	56.8%	92.6%
	Generalisation	93.4%	66.2%	91.8%	87.7%
	Simplicity	86%	65.7%	72.6%	69.2%

The process models are generated using the inductive miner infrequent. Compared to the ones created by applying the method to an optimal event log, the process models show slightly higher complexity, mainly due to the presence of additional parallel activities. However, compared to the process model representing the material flows for the entire factory, particularly when using scanning points as classifier, the value stream process models still improve in simplicity.

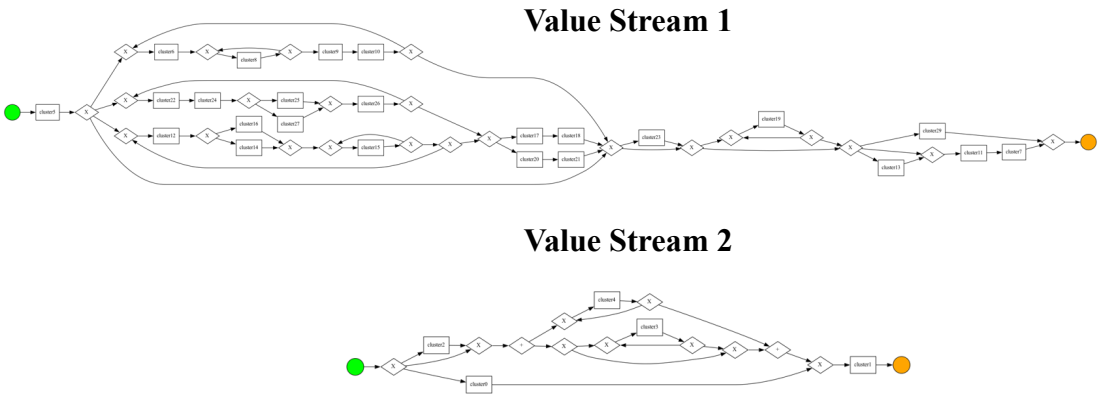


Figure 6-20: Process models (BPMN) using clustered scanning points as classifier

Based on the evaluation shown in Table 6-5, it is possible to conclude the method for pre-processing event logs is an effective solution for the identification of key value streams in cases where information about work operations is not explicitly recorded.

6.3 EVALUATING THE IMPACT OF QUALITY ISSUES

The use of a complete event log enables the simulation of different scenarios. Therefore, the event log is modified to simulate each quality problem and evaluate the effect compared to the optimal event log.

For the analysis of the individual effect of missing, incorrect, or imprecise attributes, the original event log is manipulated to simulate the different scenarios by altering 10% of the events. Depending on the implications of each quality issue, the event log is cleaned or filtered using the policies defined in Table 5-5. In some cases, a stratified sample is taken from the modified event log based on the routing frequencies, with the size of the sample being chosen depending on the size of the repaired dataset. Finally, the core method for PFA is applied and the results are evaluated by comparing the product family composition and related process models to the ones generated using an optimal event (see Figure 6-21).

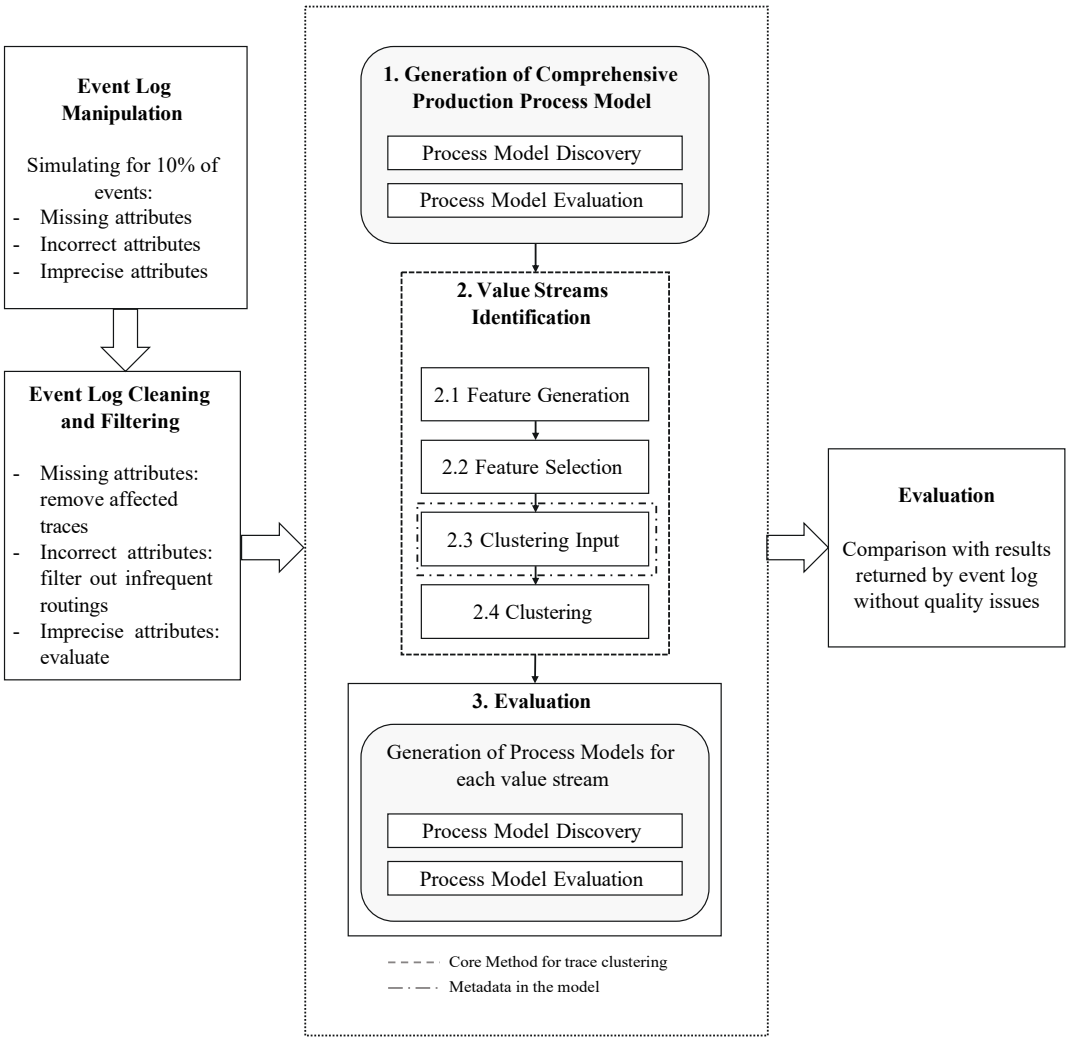


Figure 6-21: Simulation process for assessing the impact of data quality issues

6.3.1 Simulation of Quality Issues

To simulate data with missing attributes, values associated with 10% of the events are removed and replaced with missing ones.

For the scenarios simulating incorrect attributes, 10% of events are modified by changing the case or operation to another random value among the ones contained in the event log. For incorrect timestamps, 10% of the events is replaced with a random later date within the same month.

To simulate imprecise cases, 10% of the events is duplicated and the product serial associated to them (i.e. case) is changed to a random value among the ones present in the event log. For imprecise operations 8 operations have been grouped into 4 clusters with two activities each. Finally, for imprecise timestamps, the original format of the timestamps “dd/MM/yyyy hh:mm:ss” is changed to a date-only format “dd/MM/yyyy”.

6.3.2 Results

Quality issues in production data can have different effects on the proposed method for identifying and visualising value streams. For the purpose of PFA, the quality of the event log can affect results in two different ways. On one hand, it may affect the ability to identify product families. On the other hand, it may impact the quality and the interpretability of the process models.

In regard to the ability to identify product families, the results demonstrate that the method is robust in presence of quality issues affecting up to 10% of events or data coarseness. This means that the method is able to identify product families even when random data collection errors occur, or when data management infrastructures that collect coarse-grained data are used.

Table 6-7: Evaluation of event logs with quality issues and comparison with optimal event log

		Davies Bouldin Index	Average within cluster similarity	
			Value Stream 1	Value Stream 2
Optimal Event Log		0.597	96.4%	99.4%
Missing Attribute	Cases	0.597	95.7%	99.6%
	Operations	0.529	96.9%	97.9%
	Timestamps	0.543	96.7%	99.3%
Incorrect Attribute	Cases	0.546	99.5%	93.9%
	Operations	0.545	99.7%	97.9%
	Timestamps	0.535	97.7%	98.2%
Imprecise Attribute	Cases	0.560	99.7%	64.9%
	Operations	0.623	97.4%	95.8%
	Timestamps	0.464	98.9%	98.5%

Random collection errors can be caused by logging errors that affect the correctness of specific attributes. The same applies to issues related to imprecise cases, for which events may be attributed to multiple cases. This generates higher variability in the distribution of routing frequencies, as there are additional inaccurate routings recorded in the event log. Filtering out infrequent traces for each product proves to be an essential step for limiting the effect of such issues. For these scenarios, increasing the threshold for routing frequency is required to prevent the generation of inaccurate product families. Within the core method for value stream identification, feature selection is also an important step for handling the presence of incorrect values. When trace profiles are generated, a higher number of transitions is created, and feature selection can reduce the presence of such attributes.

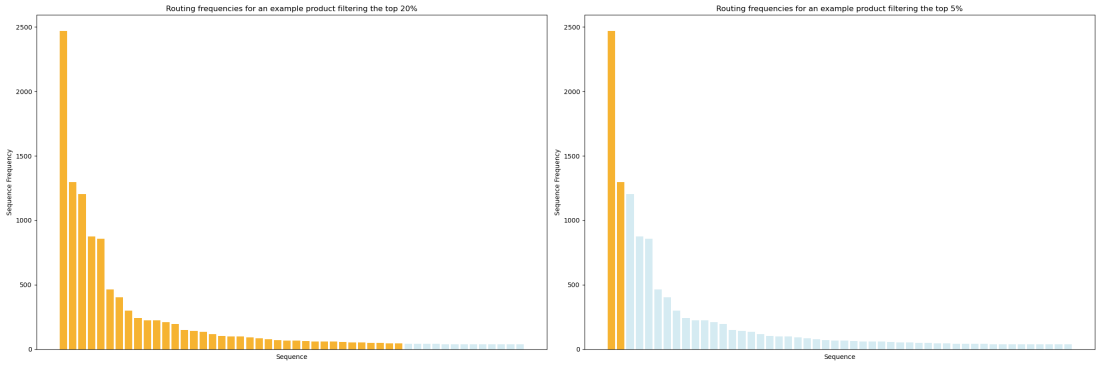


Figure 6-22: Routing frequency threshold comparison for a product in the presence of incorrect operations (20% on the left, 5% on the right)

The impact of coarse-grained data (i.e. imprecise operations and imprecise timestamps) is limited by the characteristics of the trace profiles that include both activities and transitions. The presence of imprecise operations and imprecise timestamps affects the ability to capture transitions within a sequence of operations or within the time interval of data collection (e.g. a day) respectively. However, the inclusion of activities in the trace profiles provides enough information for identifying homogeneity within routings, thus overcoming the issue of coarse-grained data. Robustness in these scenarios is important for ensuring applicability in situations where companies with limited traceability capabilities may not be able to collect enough data from their factories.

While the ability to identify product families does not get affected, the quality of process models worsens in the presence of quality issues, particularly for incorrect and imprecise attributes.

The presence of missing operations has little impact on the process models. This is because Process Mining generates models by abstracting the behaviour contained in a sample of data, and removing the traces affected by missing values does not affect the recorded behaviour. Based on which traces get affected, the model may not be able to capture some material flows, potentially displaying some minor deviations from the optimal use case.

Despite filtering out infrequent traces, the presence of incorrect attributes and imprecise cases has a small impact on the accuracy and quality of process models, particularly in precision for value stream 2.

Finally, coarse-grained data results in process models depicting inaccurate sequences. Imprecise operations are generated from situations where multiple activities are recorded as one, and they translate in process models that are less accurate and precise. The presence of imprecise timestamps, generated by situations where events are only recorded sporadically (e.g. on a daily basis), results in less precise process models with a higher number of operations occurring in parallel, as it is not possible to capture sequences occurring within the recording interval.

Table 6-8: Product family compositions using event logs with quality issues

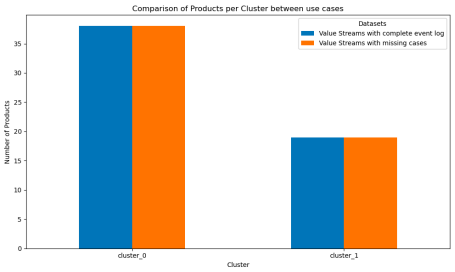
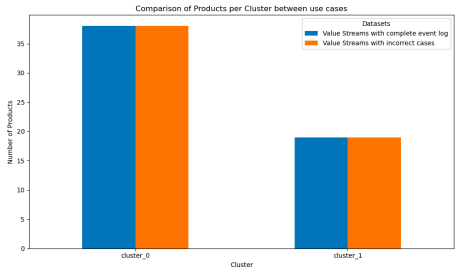
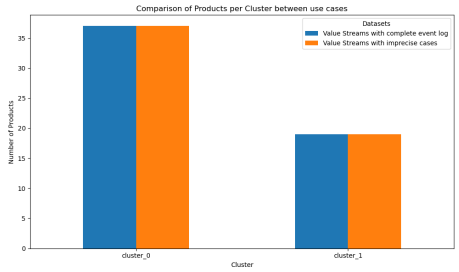
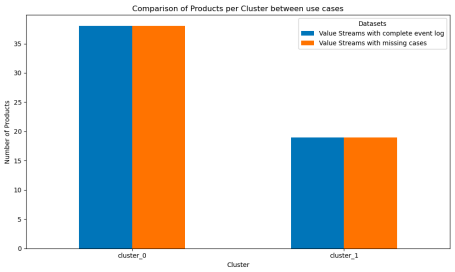
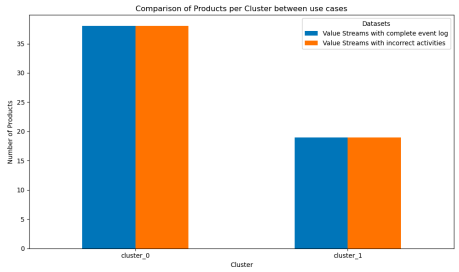
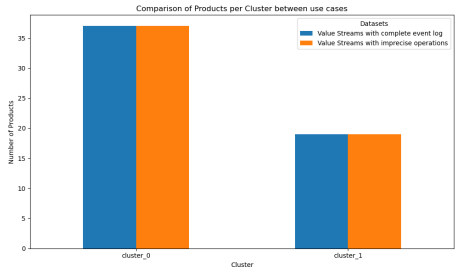
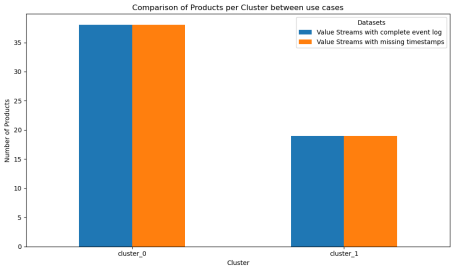
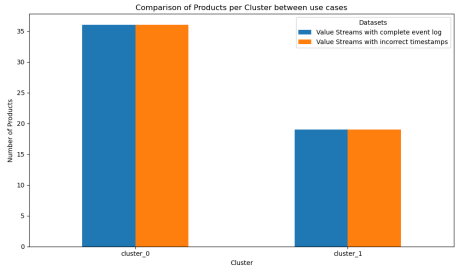
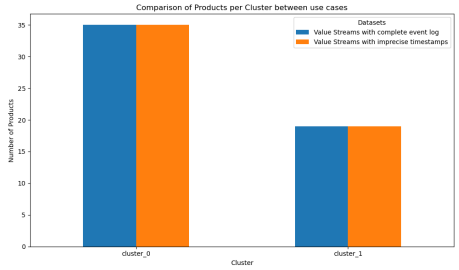
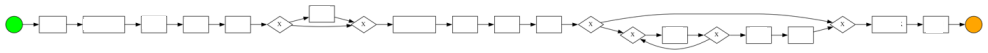
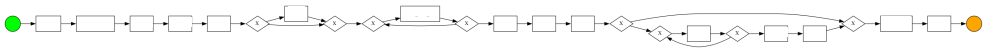
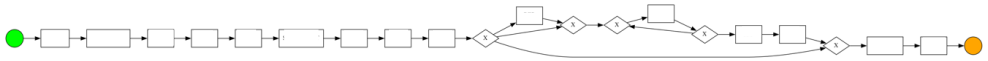
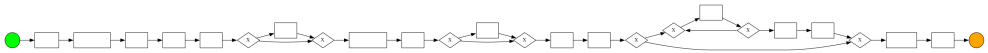
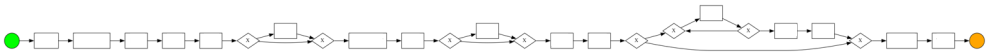
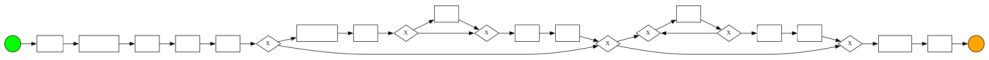
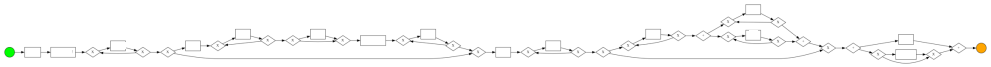
	Missing attribute	Incorrect attribute	Imprecise attribute
Case: serial number			
Activity name: operation ID			
Timestamp			

Table 6-9: Process models (BPMN) evaluation using event logs with quality issues for value stream 1

		Process Model	Replay Fitness	Precision	Generalisation	Simplicity
Complete Log			**97.8%**	**86.5%**	**93.4%**	**86%**
Missing Attribute	Cases		97.9%	85.7%	97.9%	89.5%
	Operations		98.1%	98.3%	97.9%	89.9%
	Timestamps		97.6%	81%	96.6%	86%
Incorrect Attribute	Cases		97.6%	92.3%	96.6%	86.7%
	Operations		98.2%	86.7%	96.7%	86.1%
	Timestamps		97.9%	85.3%	91.1%	83.3%

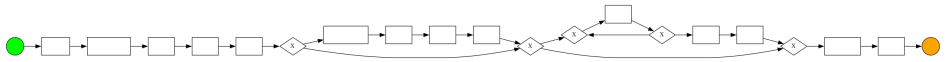
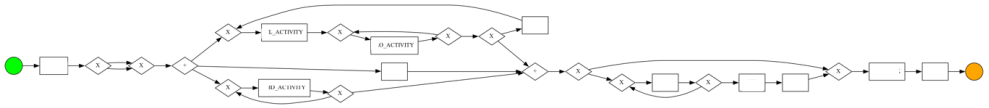
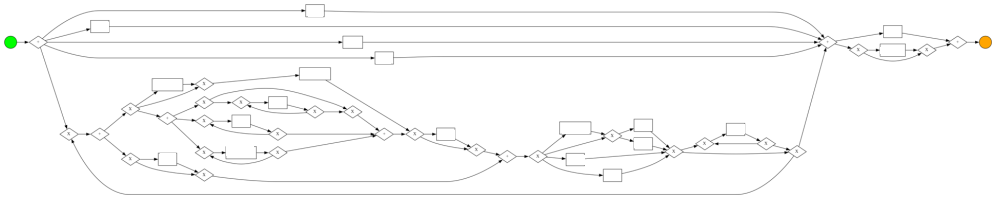
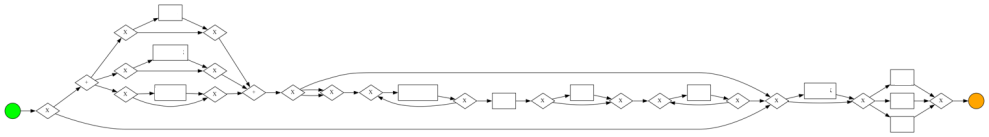
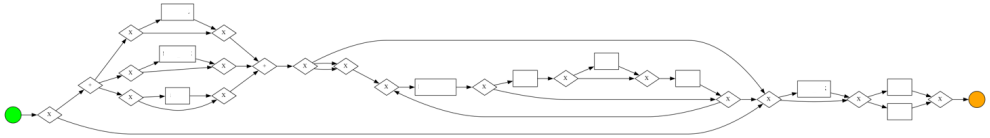
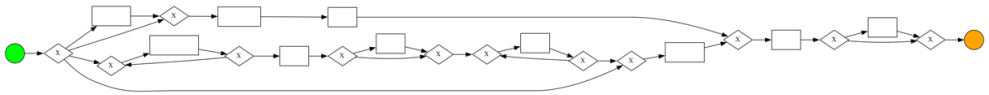
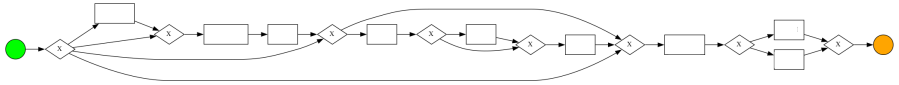
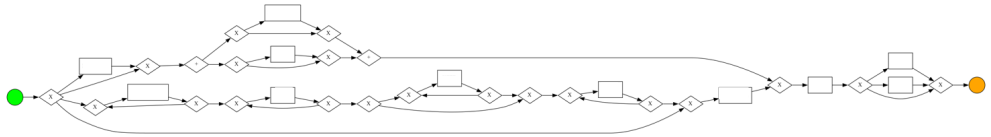
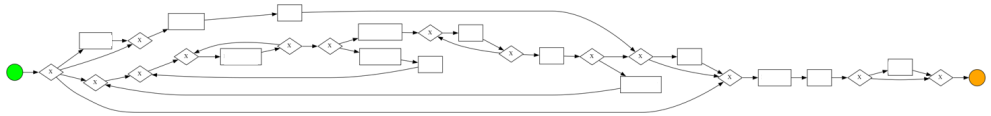
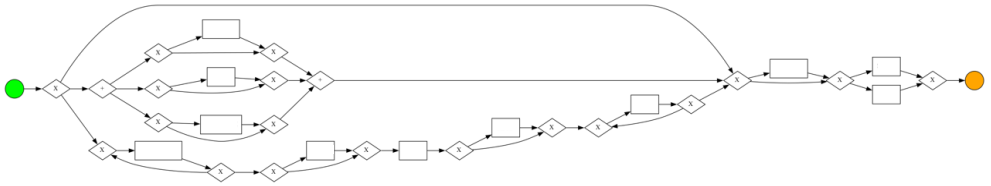
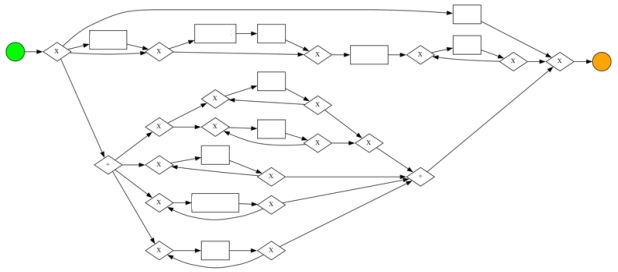
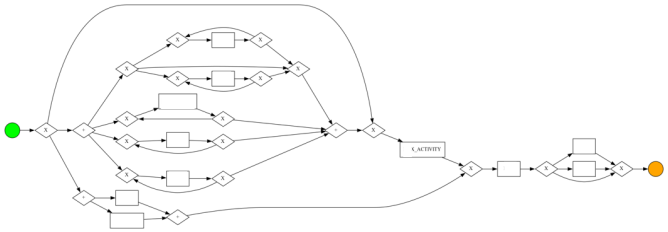
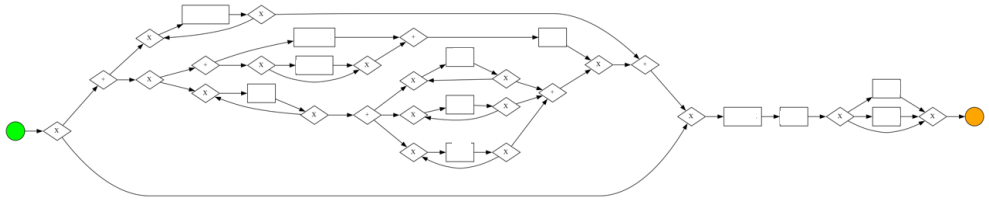
Imprecise Attribute	Cases		98.3%	85.7%	97.9%	89.5%
	Operations		98.2%	41.5%	88.9%	72.5%
	Timestamps		88.1%	23.5%	83.9%	62.6%

Table 6-10: Process models (BPMN) evaluation using event logs with quality issues for value stream 2

		Process Model	Replay Fitness	Precision	Generalisation	Simplicity
Complete Log			**91.9%**	**72.9%**	**66.2%**	**65.7%**
Missing Attribute	Cases		92%	59.2%	66.5%	67.2%
	Operations		98.5%	80.8%	78.7%	72.5%
	Timestamps		93.4%	75.4%	68.5%	67.5%

Incorrect Attribute	Cases		99.7%	83.9%	77.6%	67.3%
	Operations		99.5%	41.1%	77.4%	66.1%
	Timestamps		96%	80.8%	61.4%	64.7%
Imprecise Attribute	Cases		96%	39.2%	75.1%	65.3%

	Operations		95.8%	71.8%	71.1%	69.2%
	Timestamps		97.9%	80.4%	76.7%	64.4%

Chapter 7: Conclusions and Outlook

7.1 SUMMARY OF RESEARCH RESULTS

This research has *developed a Process Mining-based approach to identify key value streams and enable the effective application of Production Planning and Control techniques in complex environments.*

The systematic literature review provided an original synthesis of the evolution of Production Flow Analysis over three decades and the more recent developments in Process Mining applications in Manufacturing contexts. The analysis was motivated by the observation that Process Mining is a suitable technique for assisting Group Technology planning tasks (Deuse et al., 2022). Despite the potential, results from the literature review have shown that while the combination of these two techniques has received limited coverage in existing literature, Process Mining has been proven to be a useful solution for the analysis of real production processes across a variety of Manufacturing applications. As such, the review defines research gaps and by validating the research goal, it provides the theoretical groundwork for the method developed in this thesis.

While the literature review supports the theoretical foundations of this research, the industry analysis and requirements definition support its applicability. In this phase, relevant Process Mining literature has been combined with empirical data analysis to understand requirements and define possible use cases, thus informing the conceptual development of the methodology. As a result, the proposed solution takes into consideration different levels of data quality to address the challenge represented by the fact that the data capture and storage capabilities of information systems used by companies vary.

To validate the method and demonstrate its applicability, a case study is evaluated using real production data. Starting from an event log with one million events, the method was able to identify the key value streams and generate the corresponding process models, taking into account different use cases corresponding to various levels of data quality and maturity.

7.2 IMPLICATIONS AND CONTRIBUTIONS

The methodology developed in this research provides a solution for using event data for the purpose of value stream identification, which is an essential prerequisite for many Industrial Engineering improvement techniques, including Lean methods such as Value Stream Mapping, pull production, production levelling, and line balancing (Deuse et al., 2013).

From a theoretical perspective, the method developed in this research contributes to disciplines from Industrial Engineering as well as Data Science. From an Industrial Engineering perspective, this method contributes to the area of Production Planning and Control by building on the fundamental principles of Production Flow Analysis and enhancing them using Industry 4.0 techniques. While this type of synergy has received limited coverage in existing practice-oriented research (see literature review in chapter 3), a greater proportion of research has investigated the influence of Industry 4.0 on established practices in Lean Production (Buer et al., 2018; Rosin et al., 2020). In the latter case, results show that Industry 4.0 technologies can improve the implementation of Lean principles, but they do not cover their integration. Therefore, companies should continue to implement fundamental Industrial Engineering techniques such as Lean principles, while improving certain aspects using Industry 4.0 technologies (Rosin et al., 2020). The same observation can be made for Production Flow Analysis principles, as their capability to identify key value streams represents an essential enabling step for many Lean techniques and methods, including Value Stream Mapping, pull production, production levelling, and line balancing (Deuse et al., 2013). As such, the method proposed in this research demonstrates that Data Analytics techniques, namely Process Mining and Machine Learning, improve the effectiveness of Production Flow Analysis. By building on existing frameworks for trace clustering (Zandkarimi et al., 2020) and event log maturity (van der Aalst et al., 2012), this work proposes a conceptual method for value stream identification. From a Data Science perspective, the method proposed in this research demonstrates the ability to solve engineering problems related to the identification of value streams. Compared to existing research in the field of Industrial Data Science, this work provides a comprehensive methodology that takes into consideration various use cases, depending on data quality as a most relevant practical challenge. As such, while some parts of the methodology build on existing research and use established techniques, a

significant contribution is provided by the algorithm for event log pre-processing that enables the identification of the association between scanning points and operations.

From a practical perspective, this method contributes to the research enabling Industry 4.0 capabilities and foundations. Through the identification of value streams as product families and the work operations they are associated to, this method addresses some of the practical challenges of IE and enables visibility and transparency in production by facilitating the effective application of Production Planning and Control techniques.

One of the key challenges that modern manufacturing companies face is what to do with the increasing amount of data that they are able to collect from the factory floor (Kusiak, 2017). This work provides a solution to this problem, as the automated product clustering based on routings allows companies to master an unmanageable amount of data.

More broadly, the method proposed in this research demonstrates the ability to unlock visibility and transparency in production, both fundamental maturity levels of Industry 4.0 (Schuh et al., 2020). The method allows to intuitively see what happens on the factory floor by creating Digital Shadows (i.e. visibility) and apply engineering knowledge to derive and visualise key value streams (i.e. transparency). Overall, the visibility and transparency provided by this method can directly inform Production Planning and Control decisions while laying the foundations for predictive and adaptability capabilities. In terms of immediate feedback, the identification of key value streams can inform decisions related to multiple activities including production planning and scheduling, production levelling, production design, and supply chain design. In regard to the ability to unlock future capabilities, this method can serve as basis for the ability to simulate and analyse future scenarios, as representative Digital Shadows combined with engineering knowledge enable the generation of relevant recommendations and forecasts (Schuh et al., 2020). This can be taken one step further to develop Digital Twins that are able to automatically respond to changes in physical processes, thus completing the feedback loop.

Finally, in line with existing Industry 4.0 policy implications, this work represents industry-oriented research. In fact, Industry 4.0 is still at a conceptual state trying to integrate various dynamic technological concepts (F. Yang & Gu, 2021), and countries across the world are introducing different policies and national strategies to

harness the advantages of this technological shift. Since 2011, when Industry 4.0 concepts were first announced at the Hannover Fair and then included as a German strategic initiative by the government two years later (Kagermann et al., 2013), various countries followed in introducing industrial plans for enabling Industry 4.0. Some of the commonalities between the different national initiatives include a strong focus on collaboration with industry, universities, and governments, and the adoption of interdisciplinary approaches in research and development. Since 2017, as part of the Prime Minister's Industry 4.0 Taskforce, now renamed Industry 4.0 Manufacturing Forum (Gallagher, 2017), Australia has been encouraging research organisations to collaborate with industry to facilitate technological innovation in manufacturing. As a result of these national imperatives, Australian universities have adjusted the focus of their research programs, promoting PhD projects that are industry-relevant, provide 'tangible innovation' and 'identifiable impact' (Molla & Cuthbert, 2019). Other countries including Denmark, Italy, Portugal, Singapore, the United Kingdom, and the United States of America, have introduced programmes with similar goals, namely enabling to facilitate the technology transfer to the industry enabled by the collaboration between government, academia, and industry to facilitate the technology transfer to the industry (F. Yang & Gu, 2021). In line with the Industry 4.0 policy imperatives shared by countries across the world, this research contributes to the development of Industry 4.0 innovation by proposing an industry-oriented research work with 'identifiable impact'. This is achieved by proposing a method that considers multiple practical scenarios and demonstrating its applicability using real production data.

7.3 LIMITATIONS AND FUTURE WORK

This research proposes a methodology for identifying value streams and demonstrates the practical impact using an exemplary case study. In research, exemplary case studies are considered suitable methods for exploratory research investigating a contemporary set of events over which the investigator has little or no control (Yin, 2009). Multiple case experiments can lead to more generalizable conclusions by showing that findings can either be idiosyncratic to the single case study or consistently replicated (Eisenhardt & Graebner, 2007). In this research, the data used for the case study is good quality and complete, making it possible to simulate different use cases with different levels of data quality. This attempts to overcome some of the limitations of a single case study. However, exploratory research benefits from being further verified with additional experiments involving production data from other manufacturing companies. More broadly, this methodology could be adapted to processes beyond Manufacturing. While, Production Flow Analysis has been used by manufacturing companies for decades, due to its ability to objectively analyse existing material flow systems and inform restructuring planning decisions, more recent research has highlighted the potential for applying PFA to contexts beyond Manufacturing, such as service operations management (Hameri, 2011). Similarly, Process Mining techniques have demonstrated to be highly effective in discovering, analysing, and improving process across various industries, including healthcare, banking, government, education, and transportation (van der Aalst, 2016). Therefore, future research may also evaluate the applicability of the methodology proposed in this research beyond Manufacturing contexts.

This work demonstrates that the ability to identifying key value streams depends on the digital capabilities of manufacturing companies. The accuracy of results is directly correlated to the accuracy of data collection, as shown by the different use cases examined. Many manufacturers gradually introduce data collection systems by retrofitting existing equipment, as replacing old machines with limited sensory capabilities is often expensive (Lorenz et al., 2021). Therefore, in some cases the required data may simply not be captured from the physical processes, making the application of Process Mining techniques impossible. It is important to note that even in cases where data collection practices are optimal, domain expertise may be required in the interpretation of the value stream process models. In fact, the resulting process

models are the representation of the product routings, with the products being represented by unique serial numbers and a product IDs and the work operations by an activity code or a scanning point description. In practice, products may get recorded at specific check points and in some cases, these may not coincide with a specific activity. In such cases, domain knowledge is required to obtain interpretable results, and it can be applied before the event log is processed to identify value streams or after the process models have been generated. If applied in the first phases, domain knowledge requires analysing the list of unique work operations recorded in the event log and filtering down this list to the representative ones. If applied after the final results have been generated, domain knowledge requires interpreting the process models and adjusting them accordingly.

As the scope of this work is limited to the final stages of the data life-cycle (Tao et al., 2018), namely data processing, data visualisation, and data application, future research may focus on integrating this methodology with appropriate implementation architecture and technologies, thus integrating the earlier stages of the data life-cycle, from sources to data collection and storage, and addressing the limitations related to data availability. Another possible research direction is the extension of data application to enable more advanced capabilities. In fact, most Process Mining applications, including the one proposed in this research, generate Digital Shadows, that are reflections of reality automatically generated from data. While digital shadows are very useful in providing actionable insights, they do not have real-time feedback-loops, meaning that the awareness derived from the Digital Shadow does not automatically trigger actions. Instead, Digital Twins are able to trigger changes based on the insights produced (van der Aalst et al., 2021). Process Mining, and in particular object-centric Process Mining, is key in transitioning from Digital Shadows to Digital Twins (van der Aalst, 2023). As such, future research could integrate and expand the methodology developed in this work with Digital Twin applications. Similarly to the design approach used in this research, such work could be evaluated using an exemplary case study or multiple case studies.

Finally, the methodology developed in this research can be generalised to discrete production settings, where products undergo transformations individually. Potentially, the methodology can also be adapted to process manufacturing, where process models would be generated based on the movements of batches. However,

assembly production environments pose additional challenges due to merging process flows representing assembly operations of different parts identified by their own unique IDs (i.e. case IDs). Thus, identifying key value streams in these scenarios requires different Process Mining techniques. The Process Mining approaches used in this work are based on the assumption that there is a single case notation for each product, and events refer specifically to one case. Instead, assembly processes require object-centric approaches, for which it is assumed that "there are multiple case notions (called object types) and that an event may refer to any number of objects corresponding to different object types" (van der Aalst, 2019b, p. 3). Recent research has addressed the problem of single case Process Mining approaches by proposing discovery algorithms (Van Der Aalst & Berti, 2020).

Bibliography

Adriansyah, A., Munoz-Gama, J., Carmona, J., van Dongen, B. F., & van der Aalst, W. M. P. (2015). Measuring precision of modeled behavior. *Information Systems and E-Business Management*, *13*(1), 37–67. https://doi.org/10.1007/S10257-014-0234-7/FIGURES/23

Agrawal, A. K., Bhardwaj, P., & Srivastava, V. (2011). Ant colony optimization for group technology applications. *International Journal of Advanced Manufacturing Technology*, *55*(5–8), 783–795. https://doi.org/10.1007/s00170-010-3097-1

Alcácer, V., & Cruz-Machado, V. (2019). Scanning the Industry 4.0: A Literature Review on Technologies for Manufacturing Systems. In *Engineering Science and Technology, an International Journal* (Vol. 22, Issue 3, pp. 899–919). Elsevier B.V. https://doi.org/10.1016/j.jestch.2019.01.006

Aloudat, M., Kamrani, A. K., & Nasr, E. A. (2008). Cellular manufacturing performance improvement using data mining techniques. *International Journal of Knowledge Management Studies*, *2*(4), 387–405. https://doi.org/10.1504/IJKMS.2008.019748

Arikan, F., & Güngör, Z. (2005). A parametric model for cell formation and exceptional elements' problems with fuzzy parameters. *Journal of Intelligent Manufacturing*, *16*(1), 103–114. https://doi.org/10.1007/s10845-005-4827-3

Aslan, A., El-Raoui, H., Hanson, J., Vasantha, G., Quigley, J., Corney, J., & Sherlock, A. (2023). Using Worker Position Data for Human-Driven Decision Support in Labour-Intensive Manufacturing. *Sensors*, *23*(10). https://doi.org/10.3390/s23104928

Bartz-Beielstein, T., Branke, J., Mehnen, J., & Mersmann, O. (2014). Evolutionary Algorithms. *WIREs Data Mining and Knowledge Discovery*, *4*(3), 178–195. https://doi.org/10.1002/widm.1124

Berlec, T., Potočnik, P., Govekar, E., & Starbek, M. (2014). A method of production fine layout planning based on self-organising neural network clustering. *International Journal of Production Research*, *52*(24), 7209–7222. https://doi.org/10.1080/00207543.2014.910619

Berry, M., Mohamed, A., & Yap, B. W. (2019). Supervised and Unsupervised Learning for Data Science. In *Unsupervised and Semi-Supervised Learning* (Issue January). http://www.springer.com/series/15892

Berti, A., & Van Der Aalst, W. (2019). Reviving token-based replay: Increasing speed while improving diagnostics. *CEUR Workshop Proceedings*, *2371*, 87–103.

Berti, A., van Zelst, S. J., & van der Aalst, W. (2019). *Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science*. http://arxiv.org/abs/1905.06169

Berti, A., van Zelst, S., & Schuster, D. (2023). PM4Py: A process mining library for Python[Formula presented]. *Software Impacts*, *17*, 100556. https://doi.org/10.1016/j.simpa.2023.100556

Beyer, H.-G., Brucherseifer, E., Jakob, W., Pohlheim, H., Sendhoff, B., & To, T. B. (2002). *Evolutionary Algorithms - Terms and Definitions*. http://ls11-www.cs.tu-dortmund.de/people/beyer/EA-glossary/

Bhide, P., Bhandwale, A., & Kesavadas, T. (2005). Cell formation using multiple process plans. *Journal of Intelligent Manufacturing*, *16*(1), 53–65. https://doi.org/10.1007/s10845-005-4824-6

Blessing, L. T. M., & Chakrabarti, A. (2009). DRM: A Design Reseach Methodology. *DRM, a Design Research Methodology*, 13–42. https://doi.org/10.1007/978-1-84882-587-1_2

Blum, F. R. (2015). *Metrics in process discovery*.

Bohnen, F., Buhl, M., & Deuse, J. (2013). Systematic procedure for leveling of low volume and high mix production. *CIRP Journal of Manufacturing Science and Technology*, *6*(1), 53–58. https://doi.org/10.1016/J.CIRPJ.2012.10.003

Bohnen, F., Maschek, T., & Deuse, J. (2011). Leveling of low volume and high mix production based on a Group Technology approach. *CIRP Journal of Manufacturing Science and Technology*, *4*(3), 247–251. https://doi.org/10.1016/J.CIRPJ.2011.06.003

Bortolini, M., Ferrari, E., Galizia, F. G., & Regattieri, A. (2021). An optimisation model for the dynamic management of cellular reconfigurable manufacturing systems under auxiliary module availability constraints. *Journal of Manufacturing Systems*, *58*(PA), 442–451. https://doi.org/10.1016/j.jmsy.2021.01.001

Bose, R. P. J. C., Mans, R. S., & van der Aalst, W. M. P. (2013). Wanna improve process mining results? *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 127–134. https://doi.org/10.1109/CIDM.2013.6597227

Boutsinas, B. (2013). Machine-part cell formation using biclustering. *European Journal of Operational Research*, *230*(3), 563–572. https://doi.org/10.1016/j.ejor.2013.05.007

Braglia, M., Carmignani, G., & Zammori, F. (2006). A new value stream mapping approach for complex production systems. *International Journal of Production Research*, *44*(18–19), 3929–3952. https://doi.org/10.1080/00207540600690545

Buer, S.-V., Strandhagen, J. O., & 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*, *56*(8), 2924–2940. https://doi.org/10.1080/00207543.2018.1442945

Buijs, J. C. A. M., Van Dongen, B. F., & Van Der Aalst, W. M. P. (2012). On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7565 LNCS*(PART 1), 305–322. https://doi.org/10.1007/978-3-642-33606-5_19

Buijs, J. C. A. M., Van Dongen, B. F., & Van Der Aalst, W. M. P. (2014). Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity. *Https://Doi.Org/10.1142/S0218843014400012*, *23*(1). https://doi.org/10.1142/S0218843014400012

Burbidge, J. L. (1989). *Production Flow Analysis*. Oxford: Clarendon Press. https://doi.org/10.1016/0378-3804(89)90095-8

Burbidge, J. L. (1991). Production flow analysis for planning group technology. *Journal of Operations Management*, *10*(1), 5–27. https://doi.org/10.1016/0272-6963(91)90033-T

Burbidge, J. L. (1992). Change to group technology: Process organization is obsolete. *International Journal of Production Research*, *30*(5), 1209–1219. https://doi.org/10.1080/00207549208942951

Burggraf, P., Wagner, J., & Heinbach, B. (2021). Bibliometric Study on the Use of Machine Learning as Resolution Technique for Facility Layout Problems. *IEEE Access*, *9*, 22569–22586. https://doi.org/10.1109/ACCESS.2021.3054563

Burke, L., & Kamal, S. (1995). Neural networks and the part family/ machine group formation problem in cellular manufacturing: A framework using fuzzy ART. *Journal of Manufacturing Systems*, *14*(3), 148–159. https://doi.org/10.1016/0278-6125(95)98883-8

Car, Z., & Mikac, T. (2006). Evolutionary approach for solving cell-formation problem in cell manufacturing. *Advanced Engineering Informatics*, *20*(3), 227–232. https://doi.org/10.1016/j.aei.2006.01.005

Caudell, T. P. (1992). Hybrid optoelectronic adaptive resonance theory neural processor, ART1. *Applied Optics*, *31*(29), 6220. https://doi.org/10.1364/ao.31.006220

Ceylan, C., Başkurt, H., Erkan, Y., & Uğur, Ş. (2023). PROCESS ANALYSIS AND OPTIMAL FACILITY LAYOUT PLANNING IN MANUFACTURING SYSTEMS. *Yugoslav Journal of Operations Research*, *33*(1), 133–152. https://doi.org/10.2298/YJOR2105015033C

Chakraborty, K., & Roy, U. (1993). Connectionist models for part-family classification. *Computers and Industrial Engineering*, *24*(2), 189–198.

Chandra Bose, J. R., Van der Aalst, W., Jagadeesh Chandra Bose, R., Mans, R. S., & van der Aalst, W. M. (2012). *Wanna Improve Process Mining Results? It's High Time We Consider Data Quality Issues Seriously*. https://doi.org/10.1109/CIDM.2013.6597227

Chattopadhyay, M., Dan, P. K., & Mazumdar, S. (2012). Application of visual clustering properties of self organizing map in machine-part cell formation. *Applied Soft Computing Journal*, *12*(2), 600–610. https://doi.org/10.1016/j.asoc.2011.11.004

Chattopadhyay, M., Sengupta, S., Ghosh, T., Dan, P. K., & Mazumdar, S. (2013). Neuro-genetic impact on cell formation methods of Cellular Manufacturing System design: A quantitative review and analysis. *Computers and Industrial Engineering*, *64*(1), 256–272. https://doi.org/10.1016/j.cie.2012.09.016

Chen, D.-S., Chen, H.-C., & Park, J.-M. (1996). An improved ART neural net for machine cell formation. *Advanced Materials Research*, *61*, 1–6.

Chen, F. F., & Sagi, S. R. (1995a). Rdvanced manffactudng Technol ogu Concurrent Design of Manufacturing Cell and Control Functions: A Neural Network Approach. *Int J Adv Manuf Technol*, *10*, 118–130.

Chen, F. F., & Sagi, S. R. (1995b). Rdvanced manffactudng Technol ogu Concurrent Design of Manufacturing Cell and Control Functions: A Neural Network Approach. *Int J Adv Manuf Technol*, *10*, 118–130.

Chen, H. G., & Guerrero, H. H. (1994). A general search algorithm for cell formation in group technology. *International Journal of Production Research*, *32*(11), 2711–2724. https://doi.org/10.1080/00207549408957094

Chen, M. C. (2003). Configuration of cellular manufacturing systems using association rule induction. *International Journal of Production Research*, *41*(2), 381–395. https://doi.org/10.1080/0020754021000024184

Chen, M. L., Wu, C. M., & Chen, C. L. (2002). An integrated approach of art1 and tabu search to solve cell formation problems. *Journal of the Chinese Institute of Industrial Engineers*, *19*(3), 62–74. https://doi.org/10.1080/10170660209509205

Chen, S. J., & Cheng, C. S. (1994). A neural network-based cell formation algorithm in cellular manufacturing. *International Journal of Production Research*, *32*(12), 293–318. https://doi.org/10.1080/00207549508930150
Chen, S. J., & Cheng, C. S. (1995). A neural network-based cell formation algorithm in cellular manufacturing. *International Journal of Production Research*, *32*(12), 293–318. https://doi.org/10.1080/00207549508930150
Chien-Ta Bruce, C. T. B., Lenny Koh, S. C., & Kay Mahamaneerat, W. (2007). Domain-concept association rules mining for large-scale and complex cellular manufacturing tasks. *Journal of Manufacturing Technology Management*, *18*(7), 787–806. https://doi.org/10.1108/17410380710817255
Cho, M., Park, G., Song, M., Lee, J., & Kum, E. (2021). Quality-aware resource model discovery. *Applied Sciences (Switzerland)*, *11*(12). https://doi.org/10.3390/app11125730
Cho, M., Park, G., Song, M., Lee, J., Lee, B., & Kum, E. (2021). Discovery of Resource-Oriented Transition Systems for Yield Enhancement in Semiconductor Manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, *34*(1), 17–24. https://doi.org/10.1109/TSM.2020.3045686
Choueiri, A. C., & Portela Santos, E. A. (2021a). Discovery of path-attribute dependency in manufacturing environments: A process mining approach. *Journal of Manufacturing Systems*, *61*, 54–65. https://doi.org/10.1016/j.jmsy.2021.08.005
Choueiri, A. C., & Portela Santos, E. A. (2021b). Multi-product scheduling through process mining: bridging optimization and machine process intelligence. *Journal of Intelligent Manufacturing*, *32*(6), 1649–1667. https://doi.org/10.1007/s10845-021-01767-2
Choueiri, A. C., Sato, D. M. V., Scalabrin, E. E., & Santos, E. A. P. (2020). An extended model for remaining time prediction in manufacturing systems using process mining. *Journal of Manufacturing Systems*, *56*, 188–201. https://doi.org/10.1016/j.jmsy.2020.06.003
Christodoulou, M., & Gaganis, V. I. (1998). Neural networks in manufacturing cell design. *Computers in Industry*, *36*(1–2), 133–138. https://doi.org/10.1016/s0166-3615(97)00107-3
Chu, C. H. (1997). An improved neural network for manufacturing cell formation. *Decision Support Systems*, *20*(4), 279–295. https://doi.org/10.1016/S0167-9236(97)00015-8
Chu, C. H., & Chu, C. H. (1993). Manufacturing cell formation by competitive learning. *International Journal of Production Research*, *31*(4), 829–843. https://doi.org/10.1080/00207549308956760
Chung, Y., & Kusiak, A. (1994). Grouping parts with a neural network. *Journal of Manufacturing Systems*, *13*(4), 262–275. https://doi.org/10.1016/0278-6125(94)90034-5
Conforti, R., Rosa, M. La, & Hofstede, A. H. M. ter. (2017). Filtering Out Infrequent Behavior from Business Process Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, *29*(2), 300–314. https://doi.org/10.1109/TKDE.2016.2614680
Currie, K. R. (1992). An intelligent grouping algorithm for cellular manufacturing. *Computers and Industrial Engineering*, *23*(1–4), 109–112. www.journal.uta45jakarta.ac.id

Dagli, C., & Huggahalli, R. (1995). Machine-part family formation with the adaptive resonance theory paradigm. *International Journal of Production Research*, 893–913.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*(2).

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Delcoucq, L., Dupiereux-Fettweis, T., Lecron, F., & Fortemps, P. (2023). Resource and activity clustering based on a hierarchical cell formation algorithm. *Applied Intelligence*, *53*(1), 532–541. https://doi.org/10.1007/s10489-022-03457-9

Deuse, J., Konrad, B., & Bohnen, F. (2013). Renaissance of group technology: Reducing variability to match lean production prerequisites. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, *46*(9), 998–1003. https://doi.org/10.3182/20130619-3-RU-3018.00319

Deuse, J., West, N., & Syberg, M. (2022). Rediscovering scientific management. The evolution from industrial engineering to industrial data science. *International Journal of Production Management and Engineering*, *10*(1), 1–12. https://doi.org/10.4995/ijpme.2022.16617

Dobado, D., Lozano, S., Bueno, J. M., & Larrañeta, J. (2002). Cell formation using a fuzzy min-max neural network. *International Journal of Production Research*, *40*(1), 93–107. https://doi.org/10.1080/00207540110073064

dos Santos, C. F., Loures, E. de F. R., & Santos, E. A. P. (2024). A smart framework to perform a criticality analysis in industrial maintenance using combined MCDM methods and process mining techniques. *International Journal of Advanced Manufacturing Technology*. https://doi.org/10.1007/s00170-024-13193-8

Drira, A., Pierreval, H., & Hajri-Gabouj, S. (2007). Facility layout problems: A survey. *Annual Reviews in Control*, *31*(2), 255–267. https://doi.org/10.1016/j.arcontrol.2007.04.001

Duong, L. T., Travé-Massuyès, L., Subias, A., & Roa, N. B. (2021). Assessing product quality from the production process logs. *International Journal of Advanced Manufacturing Technology*, *117*(5–6), 1615–1631. https://doi.org/10.1007/s00170-021-07764-2

Durán, O., Rodriguez, N., & Consalter, L. A. (2010). Collaborative particle swarm optimization with a data mining technique for manufacturing cell design. *Expert Systems with Applications*, *37*(2), 1563–1567. https://doi.org/10.1016/j.eswa.2009.06.061

Dy, J. G., & Brodley, C. E. (2004). Feature Selection for Unsupervised Learning. In *Journal of Machine Learning Research* (Vol. 5).

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory Building From Cases: Opportunities And Challenges. *Academy of Management Journal*, *50*(1), 25–32. https://doi.org/10.5465/amj.2007.24160888

Ejsmont, K., Gladysz, B., Corti, D., Castaño, F., Mohammed, W. M., & Martinez Lastra, J. L. (2020). Towards 'Lean Industry 4.0′ – Current trends and future perspectives. *Cogent Business & Management*, *7*(1), 1781995. https://doi.org/10.1080/23311975.2020.1781995

El-Kebbe, D. A., & Danne, C. (2006). On adapting neural networks to cellular manufacturing. *ESM 2006 - 2006 European Simulation and Modelling Conference: Modelling and Simulation 2006*, 450–455.

ElMaraghy, H. A., & Gu, P. (1988). Expert parts assignment in cellular manufacturing using pattern recognition. *International Journal of Machine Tools and Manufacture*, *28*(4), 503–514. https://doi.org/10.1016/0890-6955(88)90063-6

ElMaraghy, H. A., & Gu, P. (1989). Feature based expert parts assignment in cellular manufacturing. *Journal of Manufacturing Systems*, *8*(2), 139–152. https://doi.org/10.1016/0278-6125(89)90032-0

Enke, D., Ratanapan, K., & Dagli, C. (1998). Machine-part family formation utilizing an art1 neural network implemented on a parallel neuro-computer. *Computers and Industrial Engineering*, *34*(1), 189–205. https://doi.org/10.1016/S0360-8352(97)00160-5

Enke, D., Ratanapan, K., & Dagli, C. (2000). Large machine-part family formation utilizing a parallel ART1 neural network. *Journal of Intelligent Manufacturing*, *11*(6), 591–604. https://doi.org/10.1023/A:1026508623947

Erlach, K. (2013). *Value Stream Design*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12569-0

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/J.ENGAPPAI.2022.104743

Farooqui, A., Bengtsson, K., Falkman, P., & Fabian, M. (2019). From factory floor to process models: A data gathering approach to generate, transform, and visualize manufacturing processes. *CIRP Journal of Manufacturing Science and Technology*, *24*, 6–16. https://doi.org/10.1016/j.cirpj.2018.12.002

Fletcher, S., & Islam, Z. (2018). Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems Fletcher & Islam*, *22*.

Forghani, K., & Fatemi Ghomi, S. M. T. (2019). A queuing theory-based approach to designing cellular manufacturing systems. *Scientia Iranica*, *26*(3E), 1865–1880. https://doi.org/10.24200/sci.2018.5020.1047

Friederich, J., Francis, D. P., Lazarova-Molnar, S., & Mohamed, N. (2022). A framework for data-driven digital twins for smart manufacturing. *Computers in Industry*, *136*. https://doi.org/10.1016/j.compind.2021.103586

Gallagher, S. (2017). *Industry 4.0 Testlabs in Australia - Preparing for the Future*.

Ghosh, T., Sengupta, S., Doloi, B., & Dan, P. K. (2014). AI-based techniques in cellular manufacturing systems: A chronological survey and analysis. *International Journal of Industrial and Systems Engineering*, *17*(4), 449–476. https://doi.org/10.1504/IJISE.2014.063964

Gwiazda, A., & Knosala, R. (1997). Group technology using neural nets. *Journal of Materials Processing Technology*, 181–188.

Halaska, M., & Sperka, R. (2018). Process Mining – the Enhancement of Elements Industry 4.0. *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 1–6. https://doi.org/10.1109/ICCOINS.2018.8510578

Ham, I., Goncalves, E. V., & Han, C. P. (1988). An Integrated Approach to Group Technology Part Family Data Base Design Based on Artificial Intelligence Techniques. *CIRP Annals - Manufacturing Technology*, *37*(1), 433–437. https://doi.org/10.1016/S0007-8506(07)61671-0

Hamdani, T. M., Won, J. M., Alimi, A. M., & Karray, F. (2007). Multi-objective Feature Selection with NSGA II. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), *4431 LNCS*(PART 1), 240–247. https://doi.org/10.1007/978-3-540-71618-1_27
Hameri, A. P. (2011). Production flow analysis—Cases from manufacturing and service industry. *International Journal of Production Economics*, *129*(2), 233–241. https://doi.org/10.1016/J.IJPE.2010.10.015
Harrington, E. C. (1965). The desirability function. *Industrial Quality Control*.
Heragu, S. S. (1989). Knowledge based approach to machine cell layout. *Computers and Industrial Engineering*, *17*(1–4), 37–42.
Holland, J. H. (1973). Genetic Algorithms and the Optimal Allocation of Trials. *SIAM Journal on Computing*, *2*(2), 88–105. https://doi.org/10.1137/0202009
Hopp, W. J., & Spearman, M. L. (2011). *Factory Physics*. Long Grove, Ill. : Waveland Press.
Horsthofer-Rauch, J., Guesken, S. R., Weich, J., Rauch, A., Bittner, M., Schulz, J., & Zaeh, M. F. (2024). Sustainability-integrated value stream mapping with process mining. *Production and Manufacturing Research*, *12*(1). https://doi.org/10.1080/21693277.2024.2334294
Hosseini-Nasab, H., Fereidouni, S., Fatemi Ghomi, S. M. T., & Fakhrzad, M. B. (2018). Classification of facility layout problems: a review study. *International Journal of Advanced Manufacturing Technology*, *94*(1–4), 957–977. https://doi.org/10.1007/s00170-017-0895-8
Hu, Y. C., Chen, R. S., & Tzeng, G. H. (2002). Mining fuzzy association rules for classification problems. *Computers and Industrial Engineering*, *43*(4), 735–750. https://doi.org/10.1016/S0360-8352(02)00136-5
Hüllermeier, E. (2011). Fuzzy sets in machine learning and data mining. *Applied Soft Computing Journal*, *11*(2), 1493–1505. https://doi.org/10.1016/j.asoc.2008.01.004
Jang, I., & Rhee, J. (1997). Generalized machine cell formation considering material flow and plant layout using modified self-organizing feature maps. *Computers and Industrial Engineering*, *33*, 457–460.
Josien, K., & Liao, T. W. (2000). Integrated use of fuzzy c-means and fuzzy KNN for GT part family and machine cell formation. *International Journal of Production Research*, *38*(15), 3513–3536. https://doi.org/10.1080/002075400422770
Kagermann, H., Wahlster, W., Helbig, J., Hellinger, A., Stumpf, M. A. V., Treugut, L., Blasco, J., Galloway, H., & Findeklee, U. (2013). *Recommendations for implementing the strategic initiative INDUSTRIE 4.0. Final report of the Industrie 4.0 Working Group*.
Kamal, S. (1995). Adaptive clustering algorithm for group technology: An application of the fuzzy ART neural network. *Manufacturing Research and Technology*, *24*(C), 251–282. https://doi.org/10.1016/S1572-4417(06)80045-8
Kamal, S., & Burke, L. I. (1996). FACT: A new neural network-based clustering algorithm for group technology. *International Journal of Production Research*, *34*(4), 919–946. https://doi.org/10.1080/00207549608904943
Kaniappan Chinnathai, M., & Alkan, B. (2023). A digital life-cycle management framework for sustainable smart manufacturing in energy intensive industries. *Journal of Cleaner Production*, *419*. https://doi.org/10.1016/j.jclepro.2023.138259
Kao, Y., & Moon, Y. B. (1991). A unified group technology implementation using the backpropagation learning rule of neural networks. *Computers and Industrial Engineering*, *20*(4), 425–437. https://doi.org/10.1016/0360-8352(91)90015-X

Kao, Y., & Moon, Y. B. (1995). Feature-based memory association for group technology. *International Journal of Production Research*, *36*(6), 1653–1677. https://doi.org/10.1080/002075498193219
Kao, Y., & Moon, Y. B. (1997). Part family formation by memory association. *International Journal of Advanced Manufacturing Technology*, *13*(9), 649–657. https://doi.org/10.1007/BF01350823
Kao, Y., & Moon, Y. B. (1998). Feature-based memory association for group technology. *International Journal of Production Research*, *36*(6), 1653–1677. https://doi.org/10.1080/002075498193219
Kaparthi, S., & Suresh, N. C. (1992). Machine-component cell formation in group technology: A neural network approach. *International Journal of Production Research*, *30*(6), 1353–1367. https://doi.org/10.1080/00207549208942961
Kaparthi, S., Suresh, N. C., & Cerveny, R. P. (1993). An improved neural network leader algorithm for part-machine grouping in group technology. *European Journal of Operational Research*, *69*(3), 342–356. https://doi.org/10.1016/0377-2217(93)90020-N
Karthikeyan, S., Saravanan, M., & Ganesh Kumar, S. (2016). Designing an incremental cellular manufacturing system by using a hybrid approach based on the genetic algorithm and particle swarm optimisation. *International Journal of Enterprise Network Management*, *7*(4), 322–333. https://doi.org/10.1504/IJENM.2016.080459
Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, *90*(431), 928–934. https://doi.org/10.1080/01621459.1995.10476592
Kiang, M. Y., Kulkarni, U. R., & Tam, K. Y. (1995). Self-organazing map network as interactive clustering tool - An application to group technology. *Decision Support Systems*, *15*, 351–374.
King, J. R. (1980). Machine-component grouping in production flow analysis: An approach using a rank order clustering algorithm. *International Journal of Production Research*, *18*(2), 213–232. https://doi.org/10.1080/00207548008919662
Knoll, D., Reinhart, G., & Prüglmeier, M. (2019). Enabling value stream mapping for internal logistics using multidimensional process mining. *Expert Systems with Applications*, *124*, 130–142. https://doi.org/10.1016/j.eswa.2019.01.026
Kong, T., Seong, K., Song, K., & Lee, K. (2018). Two-mode modularity clustering of parts and activities for cell formation problems. *Computers and Operations Research*, *100*, 77–88. https://doi.org/10.1016/j.cor.2018.06.018
Krajčovič, M., Bastiuchenko, V., Furmannová, B., Botka, M., & Komačka, D. (2024). New Approach to the Analysis of Manufacturing Processes with the Support of Data Science. *Processes*, *12*(3). https://doi.org/10.3390/pr12030449
Kulkarni, U. R., & Kiang, M. Y. (1995). Dynamic grouping of parts in flexible manufacturing systems - a self-organizing neural networks approach. *European Journal of Operational Research*, *84*(1), 192–212. https://doi.org/10.1016/0377-2217(94)00326-8
Kumbhar, M., Ng, A. H. C., & Bandaru, S. (2023). A digital twin based framework for detection, diagnosis, and improvement of throughput bottlenecks. *Journal of Manufacturing Systems*, *66*, 92–106. https://doi.org/10.1016/j.jmsy.2022.11.016
Kuo, R. J., Chi, S. C., & Teng, P. W. (2001). Generalized part family formation through fuzzy self-organizing feature map neural network. *Computers and*

Industrial Engineering, *40*(1–2), 79–100. https://doi.org/10.1016/S0360-8352(00)00073-5

Kuo, R. J., Su, Y. T., Chiu, C. Y., Chen, K. Y., & Tien, F. C. (2006). Part family formation through fuzzy ART2 neural network. *Decision Support Systems*, *42*(1), 89–103. https://doi.org/10.1016/j.dss.2004.10.012

Kusiak, A. (1988). EXGT-s: A knowledge based system for group technology. *International Journal of Production Research*, *26*(5), 887–904. https://doi.org/10.1080/00207548808947908

Kusiak, A. (2017). Smart manufacturing must embrace big data. *Nature 2017 544:7648*, *544*(7648), 23–25. https://doi.org/10.1038/544023a

Laghouag, A., Zafrah, F. bin, Qureshi, M. R. N. M., & Sahli, A. A. (2024). Eliminating Non-Value-Added Activities and Optimizing Manufacturing Processes Using Process Mining: A Stock of Challenges for Family SMEs. *Sustainability (Switzerland)*, *16*(4). https://doi.org/10.3390/su16041694

Lasi, H., Fettke, P., Kemper, H. G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business and Information Systems Engineering*, *6*(4), 239–242. https://doi.org/10.1007/s12599-014-0334-4

Lee, C. K. H., Choy, K. L., Ho, G. T. S., & Lam, C. H. Y. (2016). A slippery genetic algorithm-based process mining system for achieving better quality assurance in the garment industry. *Expert Systems with Applications*, *46*, 236–248. https://doi.org/10.1016/j.eswa.2015.10.035

Lee, C. K. H., Ho, G. T. S., Choy, K. L., & Pang, G. K. H. (2014). A RFID-based recursive process mining system for quality assurance in the garment industry. *International Journal of Production Research*, *52*(14), 4216–4238. https://doi.org/10.1080/00207543.2013.869632

Lee, H., Malavé, C. O., & Ramachandran, S. (1992). A self-organizing neural network approach for the design of cellular manufacturing systems. *Journal of Intelligent Manufacturing*, *3*(5), 325–332. https://doi.org/10.1007/BF01577273

Lee, S. K., Kim, B., Huh, M., Cho, S., Park, S., & Lee, D. (2013). Mining transportation logs for understanding the after-assembly block manufacturing process in the shipbuilding industry. *Expert Systems with Applications*, *40*(1), 83–95. https://doi.org/10.1016/j.eswa.2012.07.033

Lee, S. Y., & Chen, T. C. (2001). Using evolutionary computation approach to improve the performance of the fuzzy-art for grouping parts. *Journal of the Chinese Institute of Industrial Engineers*, *18*(5), 55–62. https://doi.org/10.1080/10170660109509505

Lee, S. Y., & Fischer, G. W. (1999). Grouping parts based on geometrical shapes and manufacturing attributes using a neural network. *Journal of Intelligent Manufacturing*, *10*(2), 199–209. https://doi.org/10.1023/A:1008932922695

Leemans, S. J. J., Fahland, D., & van der Aalst, W. M. (2015). Scalable process discovery with guarantees. *Enterprise, Business-Process and Information Systems Modeling: 16th International Conference, BPMDS 2015, 20th International Conference, EMMSAD 2015*, 85–101. https://www.vdaalst.com/publications/p825.pdf

Leemans, S. J. J., Fahland, D., & van der Aalst, W. M. P. (2014). Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. *Lecture Notes in Business Information Processing*, *171 171 LNBIP*, 66–78. https://doi.org/10.1007/978-3-319-06257-0_6

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science*, *9*, 181–211. https://doi.org/10.28945/479

Li, S., Xu, L. Da, & Zhao, S. (2015). The internet of things: a survey. *Information Systems Frontiers*, *17*(2), 243–259. https://doi.org/10.1007/S10796-014-9492-7/FIGURES/7

Li, X., Baki, M. F., & Aneja, Y. P. (2010). An ant colony optimization metaheuristic for machinepart cell formation problems. *Computers and Operations Research*, *37*(12), 2071–2081. https://doi.org/10.1016/j.cor.2010.02.007

Liang, M., & Zolfaghari, S. (1999). Machine cell formation considering processing times and machine capacities: An ortho-synapse Hopfield neural network approach. *Journal of Intelligent Manufacturing*, *10*(5), 437–447. https://doi.org/10.1023/A:1008923114466

Liao, T. W. (1994). Design of line-type cellular manufacturing systems for minimum operating and material-handling costs. *International Journal of Production Research*, *32*(2), 387–397. https://doi.org/10.1080/00207549408956939

Liao, T. W., & Chen, L. J. (1993). An evaluation of ART1 neural models for GT part family and machine cell forming. *Journal of Manufacturing Systems*, *12*(4), 282–290. https://doi.org/10.1016/0278-6125(93)90319-O

Liao, T. W., & Lee, K. S. (1994). Integration of a feature-based CAD system and ART1 neural model for GT coding and part family forming. *Computers and Industrial Engineering*, *26*(I), 93–104.

Liker, J. (2020). *The Toyota Way* (Second Edition). McGraw Hill.

Lorenz, R., Senoner, J., Sihn, W., & Netland, T. (2021). Using process mining to improve productivity in make-to-stock manufacturing. *International Journal of Production Research*, *59*(16), 4869–4880. https://doi.org/10.1080/00207543.2021.1906460

Lozano, S., Canca, D., Guerrero, F., & García, J. M. (2001). Machine grouping using sequence-based similarity coefficients and neural networks. *Robotics and Computer-Integrated Manufacturing*, *17*(5), 399–404. https://doi.org/10.1016/S0736-5845(01)00015-1

Lu, S. C. Y., & Ham, I. (1989). Machine Learning Techniques for Group Technology Applications. *CIRP Annals - Manufacturing Technology*, *38*(1), 455–459. https://doi.org/10.1016/S0007-8506(07)62745-0

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of Industrial Information Integration*, *6*, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Lugaresi, G., Ciappina, A. D., Rossi, M., & Matta, A. (2023). Exploiting a combined process mining approach to enhance the discovery and analysis of support processes in manufacturing. *International Journal of Computer Integrated Manufacturing*, *36*(2), 169–189. https://doi.org/10.1080/0951192X.2022.2090024

Lugaresi, G., & Matta, A. (2021). Automated manufacturing system discovery and digital twin generation. *Journal of Manufacturing Systems*, *59*(January), 51–66. https://doi.org/10.1016/j.jmsy.2021.01.005

Lugaresi, G., & Matta, A. (2023). Automated digital twin generation of manufacturing systems with complex material flows: graph model completion. *Computers in Industry*, *151*. https://doi.org/10.1016/j.compind.2023.103977

Mahdavi, I., Javadi, B., Fallah-Alipour, K., & Slomp, J. (2007). Designing a new mathematical model for cellular manufacturing system based on cell utilization.

Applied Mathematics and Computation, *190*(1), 662–670. https://doi.org/10.1016/j.amc.2007.01.060

Mahdavi, I., Kaushal, O. P., & Chandra, M. (2001). Graph-neural network approach in cellular manufacturing on the basis of a binary system. *International Journal of Production Research*, *39*(13), 2913–2922. https://doi.org/10.1080/00207540110059l4

Mahmoodian, V., Jabbarzadeh, A., Rezazadeh, H., & Barzinpour, F. (2019). A novel intelligent particle swarm optimization algorithm for solving cell formation problem. *Neural Computing and Applications*, *31*, 801–815. https://doi.org/10.1007/s00521-017-3020-x

Malavé, C. O., & Ramachandran, S. (1991). Neural network-based design of cellular manufacturing systems. *Journal of Intelligent Manufacturing*, *2*(5), 305–314. https://doi.org/10.1007/BF01471178

McCutcheon, D. M., & Meredith, J. R. (1993). Conducting case study research in operations management. *Journal of Operations Management*, *11*(3), 239–256. https://doi.org/10.1016/0272-6963(93)90002-7

Mierswa, I., & Wurst, M. (2006). Information Preserving Multi-Objective Feature Selection for Unsupervised Learning. *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms*. MIT Press.

Mitrofanov, S. (1961). *Scientific Principles of Group Technology*.

Molla, T., & Cuthbert, D. (2019). Calibrating the PhD for Industry 4.0: global concerns, national agendas and Australian institutional responses. *Policy Reviews in Higher Education*, *3*(2), 167–188. https://doi.org/10.1080/23322969.2019.1637772

Moon, J., Park, G., Yang, M., & Jeong, J. (2022). Design and Verification of Process Discovery Based on NLP Approach and Visualization for Manufacturing Industry. *Sustainability (Switzerland)*, *14*(3). https://doi.org/10.3390/su14031103

Moon, Y. B. (1990). Forming part-machine families for cellular manufacturing: A neural-network approach. *The International Journal of Advanced Manufacturing Technology*, *5*(4), 278–291. https://doi.org/10.1007/BF02601537

Moon, Y. B. (1992). Establishment of a neurocomputing model for part family/machine group identification. *Journal of Intelligent Manufacturing*, *3*(3), 173–182. https://doi.org/10.1007/BF01477600

Moon, Y. B., & Chi, S. C. (1992). Generalized part family formation using neural network techniques. *Journal of Manufacturing Systems*, *11*(3), 149–159. https://doi.org/10.1016/0278-6125(92)90001-V

Moon, Y. B., & Kao, Y. (1993). Automatic generation of group technology families during the part classification process. *International Journal of Advanced Manufacturing Technology*, *8*, 231–235.

Moon, Y. B., & Roy, U. (1992). Learning group-technology part families from solid models by parallel distributed processing. *The International Journal of Advanced Manufacturing Technology*, *7*(2), 109–118. https://doi.org/10.1007/BF02601577

Mukattash, A., Idwan, S., Ashhab, M. S., Samhouri, M., & Matar, I. A. (2011). Optimal computerized model for designing cellular manufacturing systems using neural network. *Journal of Applied Sciences*, *11*(15), 2837–2842. https://doi.org/10.3923/jas.2011.2837.2842

Müllner, D. (2011). *Modern hierarchical, agglomerative clustering algorithms*. http://arxiv.org/abs/1109.2378

Muñoz-Gama, J., & Carmona, J. (2010). *A Fresh Look at Precision in Process Conformance* (pp. 211–226). https://doi.org/10.1007/978-3-642-15618-2_16

Olsen, T. L., & Tomlin, B. (2020). Industry 4.0: Opportunities and Challenges for Operations Management. *Manufacturing & Service Operations Management*, *22*(1), 113–122. https://doi.org/10.1287/msom.2019.0796

Özdemir, R. G., Gençyilmaz, G., & Aktin, T. (2007). The modified fuzzy art and a two-stage clustering approach to cell design. *Information Sciences*, *177*(23), 5219–5236. https://doi.org/10.1016/j.ins.2007.06.027

Pai, P. F., & Lee, E. S. (2001). Parts clustering by self-organizing map neural network in a fuzzy environment. *Computers and Mathematics with Applications*, *42*(1–2), 179–188. https://doi.org/10.1016/S0898-1221(01)00142-0

Pandian, R. S., & Mahapatra, S. S. (2010). Cell formation with operational time using ART1 networks. *International Journal of Services and Operations Management*, *6*(4), 377–397. https://doi.org/10.1504/IJSOM.2010.032915

Papaioannou, G., & Wilson, J. M. (2010). The evolution of cell formation problem methodologies based on recent studies (1997-2008): Review and directions for future research. *European Journal of Operational Research*, *206*(3), 509–521. https://doi.org/10.1016/j.ejor.2009.10.020

Park, J., Lee, D., & Zhu, J. (2014). An integrated approach for ship block manufacturing process performance evaluation: Case from a Korean shipbuilding company. *International Journal of Production Economics*, *156*, 214–222. https://doi.org/10.1016/j.ijpe.2014.06.012

Park, S., & Suresh, N. C. (2003). Performance of Fuzzy ART neural network and hierarchical clustering for part-machine grouping based on operation sequences. *International Journal of Production Research*, *41*(14), 3185–3216. https://doi.org/10.1080/0020754031000110277

Pegoraro, M., & van der Aalst, W. M. P. (2019). Mining Uncertain Event Data in Process Mining. *2019 International Conference on Process Mining (ICPM)*, 89–96. https://doi.org/10.1109/ICPM.2019.00023

Peker, A., & Kara, Y. (2004). Parameter setting of the Fuzzy ART neural network to part-machine cell formation problem. *International Journal of Production Research*, *42*(6), 1257–1278. https://doi.org/10.1080/00207540310001632457

Pelleg, D., & Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. *Proceedings of the Seventeenth International Conference on Machine Learning*.

Peng, W., Zhang, Z., Hildebrant, R., & Ren, S. (2021). Empirical Studies of Three Commonly Used Process Mining Algorithms. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2492–2499. https://doi.org/10.1109/SMC52423.2021.9658861

Pilot, T., & Knosala, R. (1998). The application of neural networks in group technology. *Journal of Materials Processing Technology*, *78*(1–3), 150–155. https://doi.org/10.1016/S0924-0136(97)00477-9

Rabbani, M., Keyhanian, S., Manavizadeh, N., & Farrokhi-Asl, H. (2017). Integrated dynamic cell formation-production planning: A new mathematical model. *Scientia Iranica*, *24*(5), 2550–2566. https://doi.org/10.24200/sci.2017.4387

Rafiei, H., & Ghodsi, R. (2013). A bi-objective mathematical model toward dynamic cell formation considering labor utilization. *Applied Mathematical Modelling*, *37*(4), 2308–2316. https://doi.org/10.1016/j.apm.2012.05.015

Raguraman, T. R., Sudhakara Pandian, R., & Kamalakannan, R. (2020). Threshold algorithm for the cell formation problem. *International Journal of Advanced Intelligence Paradigms*, *16*(3–4), 368–380. https://doi.org/10.1504/IJAIP.2020.107538

Ramos-Soto, A., Dacal-Nieto, A., Martín Alcrudo, G., Mosquera, G., & Areal, J. J. (2024). Analysis of automated guided vehicles performance based on process mining techniques. *Data Technologies and Applications*, *58*(2), 280–292. https://doi.org/10.1108/DTA-02-2023-0054

Rao, H. A., & Gu, P. (1994). Expert self-organizing neural network for the design of cellular manufacturing systems. *Journal of Manufacturing Systems*, *13*(5), 346–358. https://doi.org/10.1016/0278-6125(94)P2584-2

Rao, H. A., & Gu, P. (1995). A multi-constraint neural network for the pragmatic design of cellular manufacturing systems. *International Journal of Production Research*, *33*(4), 1049–1070. https://doi.org/10.1080/00207549508930193

Reinkemeyer, L. (2024). *Process Intelligence in Action* (L. Reinkemeyer, Ed.). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61343-2

Renzi, C., Leali, F., Cavazzuti, M., & Andrisano, A. O. (2014). A review on artificial intelligence applications to the optimal design of dedicated and reconfigurable manufacturing systems. *International Journal of Advanced Manufacturing Technology*, *72*(1–4), 403–418. https://doi.org/10.1007/s00170-014-5674-1

Rezaeian, J., Javadian, N., Tavakkoli-Moghaddam, R., & Jolai, F. (2011). A hybrid approach based on the genetic algorithm and neural network to design an incremental cellular manufacturing system. *Applied Soft Computing Journal*, *11*(6), 4195–4202. https://doi.org/10.1016/j.asoc.2011.03.013

Richter, R., Syberg, M., Deuse, J., Willats, P., & Lenze, D. (2023). Creating lean value streams through proactive variability management. *International Journal of Production Research*, *61*(16), 5692–5703. https://doi.org/10.1080/00207543.2022.2111614

Rismanchian, F., & Lee, Y. H. (2017). Process Mining–Based Method of Designing and Optimizing the Layouts of Emergency Departments in Hospitals. *Health Environments Research and Design Journal*, *10*(4), 105–120. https://doi.org/10.1177/1937586716674471

Rosin, F., Forget, P., Lamouri, S., & Pellerin, R. (2020). Impacts of Industry 4.0 technologies on Lean principles. *International Journal of Production Research*, *58*(6), 1644–1661. https://doi.org/10.1080/00207543.2019.1672902

Rother, M., & Shook, J. (2003). *Learning to See value stream mapping to add value and eliminate muda*. Lean enterprise institute. http://www.lean.org/Bookstore/ProductDetails.cfm?SelectedProductId=9

Rudnitckaia, J., Venkatachalam, H. S., Essmann, R., Hruska, T., & Colombo, A. W. (2022). Screening Process Mining and Value Stream Techniques on Industrial Manufacturing Processes: Process Modelling and Bottleneck Analysis. *IEEE Access*, *10*, 24203–24214. https://doi.org/10.1109/ACCESS.2022.3152211

Ruschel, E., Rocha Loures, E. de F., & Santos, E. A. P. (2021). Performance analysis and time prediction in manufacturing systems. *Computers and Industrial Engineering*, *151*(December 2019), 106972. https://doi.org/10.1016/j.cie.2020.106972

Sagi, S. R., & Chen, F. F. (1995). A framework for intelligent design of manufacturing cells. *Journal of Intelligent Manufacturing*, *6*(3), 175–190. https://doi.org/10.1007/BF00171446

Saidi-Mehrabad, M., & Safaei, N. (2007). A new model of dynamic cell formation by a neural approach. *International Journal of Advanced Manufacturing Technology*, *33*(9–10), 1001–1009. https://doi.org/10.1007/s00170-006-0518-2

Sani, M. F., van Zelst, S. J., & van der Aalst, W. M. P. (2018). Improving Process Discovery Results by Filtering Outliers Using Conditional Behavioural Probabilities. *Lecture Notes in Business Information Processing*, *308*, 216–229. https://doi.org/10.1007/978-3-319-74030-0_16

Sarno, R., & Effendi, Y. A. (2017). Hierarchy process mining from multi-source logs. *Telkomnika (Telecommunication Computing Electronics and Control)*, *15*(4), 1960–1975. https://doi.org/10.12928/TELKOMNIKA.v15i4.6326

Schuh, G., Anderl, R., Dumitrescu, R., & Krüger, A. (2020). *acatech STUDY Industrie 4.0 Maturity Index*.

Sengupta, S., Ghosh, T., & Dan, P. K. (2011a). A hybrid neural network approach to cell formation in cellular manufacturing. *International Journal of Intelligent Systems Technologies and Applications*, *10*(4), 360–376. https://doi.org/10.1504/IJISTA.2011.045484

Sengupta, S., Ghosh, T., & Dan, P. K. (2011b). Fuzzy ART K-Means Clustering Technique: A hybrid neural network approach to cellularmanufacturing systems. *International Journal of Computer Integrated Manufacturing*, *24*(10), 927–938. https://doi.org/10.1080/0951192X.2011.602362

Sengupta, S., Ghosh, T., & Dan, P. K. (2012). A neuro-agglomerative approach to strategic design of a manufacturing cell. *International Journal of Intelligent Enterprise*, *1*(3–4), 215–232. https://doi.org/10.1504/IJIE.2012.052555

Seo, K. K., & Park, J. H. (2004). Adaptive clustering algorithm for recycling cell formation: An application of fuzzy ART neural networks. *KSME International Journal*, *18*(12), 2137–2147. https://doi.org/10.1007/BF02990218

Seo, K.-K., & Park, J.-H. (2004). Adaptive clustering algorithm for recycling cell formation: An Application of fuzzy ART neural networks. *KSME International Journal*, *18*(12), 2137–2147. https://doi.org/10.1007/BF02990218

Shtub, A. (1988). Capacity allocation and material flow in planning group technology cells. *Engineering Costs and Production Economics*, *13*(3), 217–228. https://doi.org/10.1016/0167-188X(88)90008-0

Singgih, I. K. (2021). Production flow analysis in a semiconductor fab using machine learning techniques. *Processes*, *9*(3), 1–18. https://doi.org/10.3390/pr9030407

Sivasankaran, P., & Shahabudeen, P. (2014). Literature review of assembly line balancing problems. *International Journal of Advanced Manufacturing Technology*, *73*, 1665–1694. https://doi.org/10.1007/s00170-014-5944-y

Slomp, J. ;, Bokhorst, J. A. C. ;, Germs, R., Slomp, J., Bokhorst, J. A. C., & Germs, R. (2009). A lean production control system for high-variety/low-volume environments: a case study implementation. *Production Planning & Control*, *20*(7), 586–595. https://doi.org/10.1080/09537280903086164

Smalley, A. (2004). *Creating Level Pull*. Lean Enterprise Institute.

Solimanpur, M., Saeedi, S., & Mahdavi, I. (2010). Solving cell formation problem in cellular manufacturing using ant-colony-based optimization. *International Journal of Advanced Manufacturing Technology*, *50*(9–12), 1135–1144. https://doi.org/10.1007/s00170-010-2587-5

Solimanpur, M., Vrat, P., & Shankar, R. (2004). Feasibility and robustness of transiently chaotic neural networks applied to the cell formation problem. *International Journal of Production Research*, *42*(6), 1065–1082. https://doi.org/10.1080/00207543.2004.10750072

Song, M., Günther, C. W., & Van Der Aalst, W. M. P. (2009). Trace Clustering in Process Mining. *Business Process Management Workshops: BPM 2008 International Workshops, Milano, Italy, September 1-4, 2008.* .

Stanescu, A., Mata-Toledo, R. A., & Gupta, P. (2018). *Machine Learning*. Access Science. https://doi-org.ezproxy.lib.uts.edu.au/10.1036/1097-8542.395250

Sudhakara Pandian, R., & Mahapatra, S. S. (2009). Manufacturing cell formation with production data using neural networks. *Computers and Industrial Engineering*, *56*(4), 1340–1347. https://doi.org/10.1016/j.cie.2008.08.003

SudhakaraPandian, R., & Mahapatra, S. S. (2008). Cell formation with ordinal-level data using ART1-based neural networks. *International Journal of Services and Operations Management*, *5*(4), 548–573. https://econpapers.repec.org/article/idsijsoma/v_3a5_3ay_3a2009_3ai_3a4_3ap_3a548-573.htm

Suresh, N. C., & Kaparthi, S. (1994). Performance of Fuzzy ART neural network for group technology cell formation. *International Journal of Production Research*, *32*(7), 1693–1713. https://doi.org/10.1080/00207549408957030

Suresh, N. C., Slomp, J., & Kaparthi, S. (1995). The capacitated cell formation problem: A new hierarchical methodology. *International Journal of Production Research*, *33*(6), 1761–1784. https://doi.org/10.1080/00207549508930241

Suresh, N. C., Slomp, J., & Kaparthi, S. (1999). Sequence-dependent clustering of parts and machines: A Fuzzy ART neural network approach. *International Journal of Production Research*, *37*(12), 2793–2816. https://doi.org/10.1080/002075499190527

Syswerda, G. (1989). Uniform crossover in genetic algorithms. *ICGA*, *3*(2–9).

Tambuskar, D. P., Narkhede, B. E., & Mahapatra, S. S. (2018). A flexible clustering approach for virtual cell formation considering real-life production factors using Kohonen self-organising map. *International Journal of Industrial and Systems Engineering*, *28*(2), 193–215. https://doi.org/10.1504/IJISE.2018.089137

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, *48*, 157–169. https://doi.org/10.1016/j.jmsy.2018.01.006

Tao, F., Zhang, M., & Nee, A. Y. C. (2019). Background and Concept of Digital Twin. In *Digital Twin Driven Smart Manufacturing* (pp. 3–28). Elsevier. https://doi.org/10.1016/B978-0-12-817630-6.00001-1

Tortorella, G. L., Pradhan, N., Macias de Anda, E., Trevino Martinez, S., Sawhney, R., & Kumar, M. (2020). Designing lean value streams in the fourth industrial revolution era: proposition of technology-integrated guidelines. *International Journal of Production Research*, *58*(16), 5020–5033. https://doi.org/10.1080/00207543.2020.1743893

Tran, T. A., Ruppert, T., & Abonyi, J. (2021). Indoor positioning systems can revolutionise digital lean. *Applied Sciences (Switzerland)*, *11*(11). https://doi.org/10.3390/app11115291

van der Aalst, W. (2012). Process Mining: Overview and Opportunities. *ACM Transactions on Management Information Systems*, *99*(99). https://doi.org/10.1145/0000000.0000000

van der Aalst, W. (2016). Process mining: Data science in action. In *Process Mining: Data Science in Action*. https://doi.org/10.1007/978-3-662-49851-4

van der Aalst, W. (2019a). A practitioner's guide to process mining: Limitations of the directly-follows graph. *Procedia Computer Science*, *164*, 321–328. https://doi.org/10.1016/J.PROCS.2019.12.189

van der Aalst, W. (2019b). Object-Centric Process Mining: Dealing with Divergence and Convergence in Event Data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11724 LNCS*, 3–25. https://doi.org/10.1007/978-3-030-30446-1_1

van der Aalst, W. (2022). Process Mining: A 360 Degree Overview. *Lecture Notes in Business Information Processing*, 3–34. https://doi.org/10.1007/978-3-031-08848-3_1

van der Aalst, W. (2023). Toward More Realistic Simulation Models Using Object-Centric Process Mining. *ECMS 2023 Proceedings Edited by Enrico Vicario, Romeo Bandinelli, Virginia Fani, Michele Mastroianni*, 5–13. https://doi.org/10.7148/2023-0005

van der Aalst, W., Adriansyah, A., Alves De Medeiros, A. K., Arcieri, F., Baier, T., Blickle, T., Chandra Bose, J., Van Den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., De Leoni, M., … Wynn, M. (2012). Process Mining Manifesto. *Lecture Notes in Business Information Processing*. https://doi.org/https://doi.org/10.1007/978-3-642-28108-2_19

van der Aalst, W., & Berti, A. (2020). Discovering Object-centric Petri Nets. *Fundamenta Informaticae*, *175*(1–4), 1–40. https://doi.org/10.3233/FI-2020-1946

van der Aalst, W., Hinz, O., & Weinhardt, C. (2021). Resilient Digital Twins. *Business & Information Systems Engineering*, *63*(6), 615–619. https://doi.org/10.1007/s12599-021-00721-z

van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142. https://doi.org/10.1109/TKDE.2004.47

Varela, L., Araújo, A., Ávila, P., Castro, H., & Putnik, G. (2019). Evaluation of the Relation between Lean Manufacturing, Industry 4.0, and Sustainability. *Sustainability 2019, Vol. 11, Page 1439*, *11*(5), 1439. https://doi.org/10.3390/SU11051439

Velmurugan, T., & Santhanam, T. (2011). A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach. *Information Technology Journal*, *10*(3), 478–484. https://doi.org/10.3923/itj.2011.478.484

Venkumar, P., & Haq, A. N. (2006). Fractional cell formation in group technology using modified ART1 neural networks. *International Journal of Advanced Manufacturing Technology*, *28*(7–8), 761–765. https://doi.org/10.1007/s00170-004-2421-z

Venugopal, V. (1999). Soft-computing-based approaches to the group technology problem: A state-of-the-art review. *International Journal of Production Research*, *37*(14), 3335–3357. https://doi.org/10.1080/002075499190310

Venugopal, V., & Narendran, T. T. (1994). Machine-cell formation through neural network models. *International Journal of Production Research*, *32*(9), 2105–2116. https://doi.org/10.1080/00207549408957061

Verma, S., Pant, M., & Snasel, V. (2021). A Comprehensive Review on NSGA-II for Multi-Objective Combinatorial Optimization Problems. *IEEE Access*, *9*, 57757–57791. https://doi.org/10.1109/ACCESS.2021.3070634

Vosniakos, G. C., Tsifakis, A., & Benardos, P. (2006). Neural network simulation metamodels and genetic algorithms in analysis and design of manufacturing cells. *International Journal of Advanced Manufacturing Technology*, *29*(5), 541–550. https://doi.org/10.1007/s00170-005-2535-y

Wang, J., Song, S., Lin, X., Zhu, X., & Pei, J. (2015). Cleaning structured event logs: A graph repair approach. *Proceedings - International Conference on Data Engineering*, *2015-May*, 30–41. https://doi.org/10.1109/ICDE.2015.7113270

Weijters, A. J. M. M., & Ribeiro, J. T. S. (2011). Flexible heuristics miner (FHM). *IEEE SSCI 2011: Symposium Series on Computational Intelligence - CIDM 2011: 2011 IEEE Symposium on Computational Intelligence and Data Mining*, 310–317. https://doi.org/10.1109/CIDM.2011.5949453

Womack, J. P., & Jones, D. T. (1997). Lean thinking−banish waste and create wealth in your corporation. In *Journal of the Operational Research Society* (Vol. 48, Issue 11). https://doi.org/10.1057/palgrave.jors.2600967

Won, Y., & Currie, K. R. (2007). Fuzzy ART/RRR-RSS: A two-phase neural network algorithm for part-machine grouping in cellular manufacturing. *International Journal of Production Research*, *45*(9), 2073–2104. https://doi.org/10.1080/00207540600635227

Wu, T., Li, J., Bao, J., Liu, Q., Jin, Z., & Gao, J. (2024). CarbonKG: Industrial Carbon Emission Knowledge Graph-Based Modeling and Application for Carbon Traceability of Complex Manufacturing Process. *Journal of Computing and Information Science in Engineering*, *24*(8). https://doi.org/10.1115/1.4065166

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Xu, L. Da, Xu, E. L., & Li, L. (2018). Industry 4.0: state of the art and future trends. *International Journal of Production Research*, *56*(8), 2941–2962. https://doi.org/10.1080/00207543.2018.1444806

Yadav, R., Roopa, Y. M., Lavanya, M., Ramesh, J. V. N., Chitra, N. T., & Babu, G. R. (2023). Smart Production and Manufacturing System Using Digital Twin Technology and Machine Learning. *SN Computer Science*, *4*(5). https://doi.org/10.1007/s42979-023-01976-x

Yang, F., & Gu, S. (2021). Industry 4.0, a revolution that requires technology and national strategies. *Complex & Intelligent Systems*, *7*(3), 1311–1325. https://doi.org/10.1007/s40747-020-00267-9

Yang, M. S., & Yang, J. H. (2008). Machine-part cell formation in group technology using a modified ART1 method. *European Journal of Operational Research*, *188*(1), 140–152. https://doi.org/10.1016/j.ejor.2007.03.047

Yang, X.-S. (2014). Multi-Objective Optimization. In *Nature-Inspired Optimization Algorithms* (pp. 197–211). Elsevier. https://doi.org/10.1016/B978-0-12-416743-8.00014-2

Yin, R. K. (2009). *Case study research: Design and methods: Vol. Vol. 5*. sage.

YounesSinaki, R., Sadeghi, A., Mosadegh, H., Almasarwah, N., & Suer, G. (2023). Cellular manufacturing design 1996–2021: a review and introduction to applications of Industry 4.0. *International Journal of Production Research*, *61*(16), 5585–5636. https://doi.org/10.1080/00207543.2022.2105763

Yuguang, Z., Kai, X., & Dongyan, S. (2014). Clustering and group selection of interim product in shipbuilding. *Journal of Intelligent Manufacturing*, *25*(6), 1393–1401. https://doi.org/10.1007/s10845-013-0737-y

Zandkarimi, F., Rehse, J. R., Soudmand, P., & Hoehle, H. (2020). A generic framework for trace clustering in process mining. *Proceedings - 2020 2nd International Conference on Process Mining, ICPM 2020*, 177–184. https://doi.org/10.1109/ICPM49681.2020.00034

Zeidi, J. R., Javadian, N., Tavakkoli-Moghaddam, R., & Jolai, F. (2013). A hybrid multi-objective approach based on the genetic algorithm and neural network to design an incremental cellular manufacturing system. *Computers and Industrial Engineering*, *66*(4), 1004–1014. https://doi.org/10.1016/j.cie.2013.08.015

Zeller, V., Hocken, C., Stich, V., & Volker, S. (2018). *acatech Industrie 4.0 Maturity Index-A Multidimensional Maturity Model*. 105–113. https://doi.org/10.1007/978-3-319-99707-0_14ï

Zolfaghari, S., & Liang, M. (1998). Machine cell/part family formation considering processing times and machine capacities: A simulated annealing approach. *Computers and Industrial Engineering*, *34*(4), 813–823. https://doi.org/10.1016/s0360-8352(98)00112-0

Appendices

Appendix A

Intelligent Approaches for Production Flow Analysis

Table 7-1: Methods for Intelligent Production Flow Analysis

	Machine Learning								Metaheuristics						
	Supervised Learning		Unsupervised Learning												
			ANN												
	ANN	Other	SOM	ART1	ART2	Fuzzy ART	Other	Other	Genetic Algorithm	Ant Colony	Particle Swarm Optimization	Other	Mathematical Programming	Data Mining	Process Mining
(Delcoucq et al., 2023)															x
(Bortolini et al., 2021)													x		
(Raguraman et al., 2020)												x			
(Mahmoodian et al., 2019)											x				
(Forghani & Fatemi Ghomi, 2019)													x		
(Kong et al., 2018)															x
(Tambuskar et al., 2018)			x												
(Rabbani et al., 2017)										x					
(Karthikeyan et al., 2016)									x						
(Yuguang et al., 2014)					x										
(Berlec et al., 2014)			x								x				
(Boutsinas, 2013)												x			
(Zeidi et al., 2013)	x														
(Rafiei & Ghodsi, 2013)									x	x					
(Chattopadhyay et al., 2012)			x												
(Sengupta et al., 2012)								x							
(Rezaeian et al., 2011)	x								x						
(Mukattash et al., 2011)				x											
(Agrawal et al., 2011)										x					
(Sengupta et al., 2011b)							x								
(Sengupta et al., 2011a)							x								

(X. Li et al., 2010)										x					
(Solimanpur et al., 2010)										x					
(Durán et al., 2010)											x				
(Pandian & Mahapatra, 2010)				x											
(Sudhakara Pandian & Mahapatra, 2009)				x											
(M. S. Yang & Yang, 2008)				x											
(Aloudat et al., 2008)														x	
(SudhakaraPandian & Mahapatra, 2008)				x											
(Özdemir et al., 2007)						x									
(Chien-Ta Bruce et al., 2007)															
(Saidi-Mehrabad & Safaei, 2007)													x		
(Mahdavi et al., 2007)													x		
(Won & Currie, 2007)						x									
(Kuo et al., 2006)						x									
(Car & Mikac, 2006)									x						
(Vosniakos et al., 2006)	x														
(Venkumar & Haq, 2006)				x											
(Bhide et al., 2005)												x			
(Arikan & Güngör, 2005)													x		
(Peker & Kara, 2004)						x									
(K. K. Seo & Park, 2004)						x									
(Solimanpur et al., 2004)							x								
(S. Park & Suresh, 2003)						x									
(M. C. Chen, 2003)														x	
(Hu et al., 2002)														x	
(Dobado et al., 2002)						x									
(M. L. Chen et al., 2002)						x									
(Lozano et al., 2001)	x														
(Pai & Lee, 2001)			x												
(S. Y. Lee & Chen, 2001)						x									
(Mahdavi et al., 2001)				x											
(Kuo et al., 2001)							x								
(Josien & Liao, 2000)		x													
(Enke et al., 2000)				x											
(Liang & Zolfaghari, 1999)	x														
(S. Y. Lee & Fischer, 1999)						x									
(Suresh et al., 1999)						x									
(Pilot & Knosala, 1998)			x												
(Christodoulou & Gaganis, 1998)	x														

(Zolfaghari & Liang, 1998)											x				
(Kao & Moon, 1998)							x								
(Enke et al., 1998)				x											
(Kao & Moon, 1997)	x														
(Jang & Rhee, 1997)			x												
(Chu, 1997)			x												
(Gwiazda & Knosala, 1997)			x												
(D.-S. Chen et al., 1996)				x											
(Kamal & Burke, 1996)						x									
(Kulkarni & Kiang, 1995)			x												
(Sagi & Chen, 1995)	x														
(F. F. Chen & Sagi, 1995b)	x														
(Kamal, 1995)						x									
(Burke & Kamal, 1995)						x									
(Kiang et al., 1995)			x												
(Kao & Moon, 1995)							x								
(Suresh et al., 1995)						x									
(S. J. Chen & Cheng, 1995)				x											
(Rao & Gu, 1995)	x														
(Dagli & Huggahalli, 1995)				x											
(S. J. Chen & Cheng, 1994)				x											
(Chung & Kusiak, 1994)	x														
(Rao & Gu, 1994)			x												
(Suresh & Kaparthi, 1994)				x											
(Liao, 1994)				x											
(Liao & Lee, 1994)				x											
(H. G. Chen & Guerrero, 1994)											x				
(Venugopal & Narendran, 1994)			x	x											
(Kaparthi et al., 1993)				x											
(Y. B. Moon & Kao, 1993)				x											
(Liao & Chen, 1993)				x											
(Chakraborty & Roy, 1993)			x												
(Chu & Chu, 1993)							x								
(Caudell, 1992)				x											
(H. Lee et al., 1992)			x												
(Y. B. Moon, 1992)	x														
(Y. B. Moon & Roy, 1992)	x														
(Y. B. Moon & Chi, 1992)	x														
(Currie, 1992)	x														

(Kaparthi & Suresh, 1992)				x											
(Malavé & Ramachandran, 1991)	x														
(Kao & Moon, 1991)	x														
(Y. B. Moon, 1990)							x								
(S. C. Y. Lu & Ham, 1989)								x							
(ElMaraghy & Gu, 1989)		x													
(Heragu, 1989)													x		
(ElMaraghy & Gu, 1988)		x													
(Ham et al., 1988)		x													
(Kusiak, 1988)													x		
(Shtub, 1988)													x		

Appendix B

Process Mining in Manufacturing

Table 7-2: Process Mining Applications in Manufacturing

Reference	Process Mining Type	Application
(Ramos-Soto et al., 2024)	Performance analysis	Performance analysis of Automated Guided Vehicles (AGVs)
(Krajčovič et al., 2024)	Performance analysis	General procedure for the analysis of manufacturing processes
(Laghouag et al., 2024)	Performance analysis	Process value optimisation in family SMEs
(Horsthofer-Rauch et al., 2024)	Performance analysis	Sustainability-integrated VSM
(dos Santos et al., 2024)	Performance analysis	Method for enabling risk and criticality analysis of machines and support maintenance planning
(Aslan et al., 2023)	Performance analysis	Methodology for identifying opportunities to improve capacity allocation decisions
(Kumbhar et al., 2023)	Performance analysis	Digital twin framework for bottleneck analysis
(Ceylan et al., 2023)	Performance analysis	Process analysis and layout optimisation
(Rudnitckaia et al., 2022)	Performance analysis	Approach for process modelling and bottleneck analysis
(Duong et al., 2021)	Performance analysis	Framework for assessing product quality
(Choueiri & Portela Santos, 2021a)	Performance analysis	Framework for multi-level scheduling

(Cho, Park, Song, Lee, & Kum, 2021)	Performance analysis	Quality-aware resource model mining algorithm
(Cho, Park, Song, Lee, Lee, et al., 2021)	Performance analysis	Methodology for discovering a resource-oriented transition system
(Lugaresi & Matta, 2021)	Performance analysis	Automated model generation of a manufacturing system for perfonmance estimation
(Tran et al., 2021)	Performance analysis	Dynamic Value Stream Mapping
(Lorenz et al., 2021)	Performance analysis	Procedure to improve productivity in make-to-stock manufacturing
(Knoll et al., 2019)	Performance analysis	Value stream mapping for internal logistics
(Rismanchian & Lee, 2017)	Performance analysis	Layout design and optimisation
(C. K. H. Lee et al., 2016)	Performance analysis	Intelligent system to support quality assurance in the garment manufacturing
(J. Park et al., 2014)	Performance analysis	Block manufacturing process evaluation in the shipbuilding industry
(C. K. H. Lee et al., 2014)	Performance analysis	Intelligent system to support quality assurance in the garment manufacturing
(S. K. Lee et al., 2013)	Performance analysis	Block manufacturing process evaluation in the shipbuilding industry
(Yadav et al., 2023)	Process discovery	Framework for automating the creation of simulation models for Digital Twins
(Lugaresi & Matta, 2023)	Process discovery	Method for modelling manufacturing systems with complex material flows
(Lugaresi et al., 2023)	Process discovery	Approach for discovery and analysis of production support processes
(Delcoucq et al., 2023)	Process discovery	Hierarchical cell formation approach

(Friederich et al., 2022)	Process discovery	Automated generation of simulation models for Digital Twins
(J. Moon et al., 2022)	Process discovery	Approach for automated process discovery based on NLP
(Choueiri & Portela Santos, 2021b)	Process discovery	Framework for the discovery of path-attribute dependency
(Farooqui et al., 2019)	Process discovery	Method for data generation and creation of process models describing the behaviour of manufacturing stations
(Kong et al., 2018)	Process discovery	Two-mode modularity clustering for cell formation problems
(Sarno & Effendi, 2017)	Process discovery	Process modelling using data from distributed departments
(Wu et al., 2024)	Predictive PM	Tracking carbon emission flows produced by manufacturing processes
(Kaniappan Chinnathai & Alkan, 2023)	Predictive PM	Framework for supporting sustainability and smart manufacturing in energy intensive industries
(Ruschel et al., 2021)	Predictive PM	Completion time prediction and performance analysis in manufacturing
(Choueiri et al., 2020)	Predictive PM	Model for time remaining prediction in manufacturing systems

Appendix C

Process Models Generated by the Heuristics Miner Discovery Algorithm

Table 7-3: Process Models generated by the heuristic miner algorithm for an optimal event log and after event log pre-processing

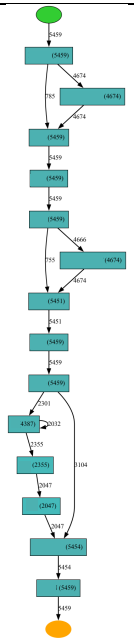
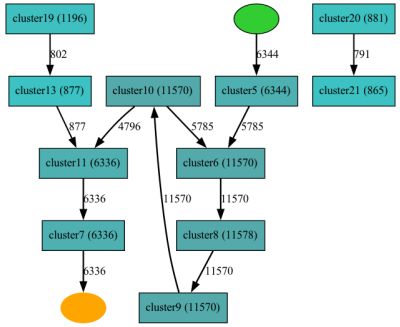
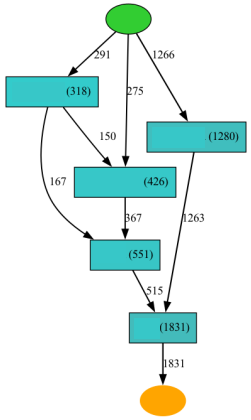
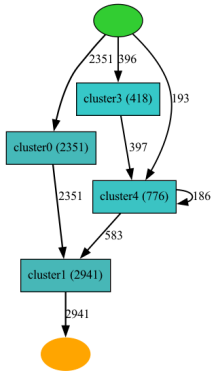
Value Stream 1		Value Stream 2	
Optimal Event Log	**Event Log Pre-Processing**	**Optimal Event Log**	**Event Log Pre-Processing**
			

Table 7-4: Process Models generated by the heuristic miner algorithm using an event log with missing attributes

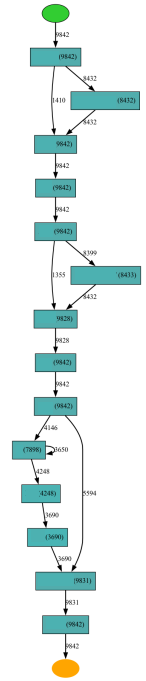
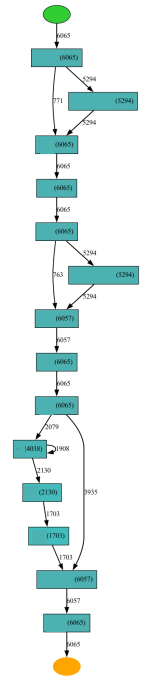
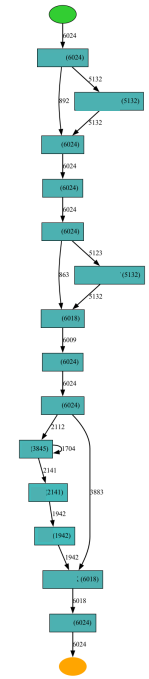
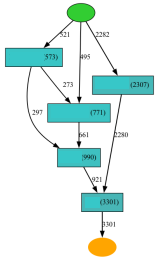
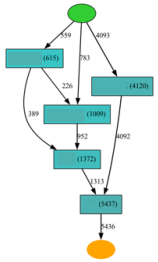
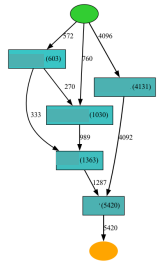
	Missing Cases	Missing Operations	Missing timestamps
Value Stream 1			
Value Stream 2			

Table 7-5: Process Models generated by the heuristic miner algorithm using an event log with incorrect attributes

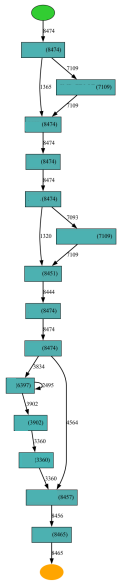
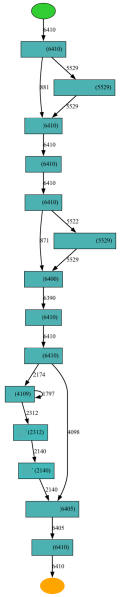
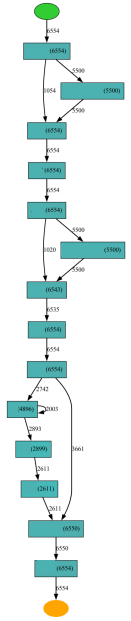
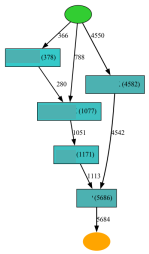
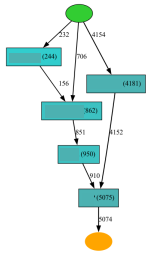
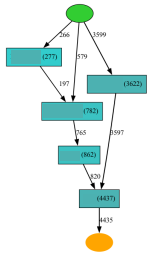
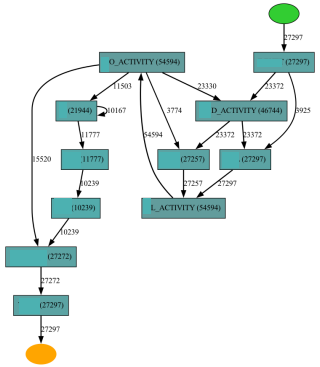
	Incorrect Cases	Incorrect Operations	Incorrect timestamps
Value Stream 1			
Value Stream 2			

Table 7-6: Process Models generated by the heuristic miner algorithm using an event log with imprecise attributes

	Imprecise Cases	Imprecise Operations	Imprecise timestamps
Value Stream 1			
Value Stream 2