Production Analytics for a novel additive manufacturing system

Rotari, Marta

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

PhD Thesis
Doctor of Philosophy

DTU Compute
Department of Applied Mathematics and Computer Science

Production Analytics for a novel additive manufacturing system

Marta Rotari

Kongens Lyngby, 2023

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Matematiktorvet
Building 303B
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary

Additive manufacturing, also known as 3D printing, has emerged as a promising solution in the context of technological advancements and innovative production processes. This thesis focuses on two additive manufacturing processes, namely Selective Laser Sintering (SLS) and Selective Thermoplastic Electrophotographic Process (STEP), which have the potential to revolutionize industrial production. However, their current status requires extensive research and development efforts to upscale them for widespread application.

Several key challenges need to be addressed to achieve the industrial-scale implementation of 3D printing. Firstly, the understanding of the 3D printing process is limited, necessitating the expansion of the engineering knowledge base, its functioning and underlying mechanisms. Secondly, optimizing the entire 3D printing process is crucial for ensuring reproducibility and consistently high-quality output. Lastly, the inherent instability of the 3D printing process poses a significant challenge to its integration into large-scale production. Three main sources of data can be identified: input data (machine parameters), process data (data acquired during the production process) and output data (the quality of the final products). During the PhD journey, some methodologies have been developed to address the challenges faced by the new 3D printers.

The Ph.D. thesis begins with an introductory chapter that discusses the context of the Ph.D. project as well as its contributions. A brief introduction to additive manufacturing, its limitations, and the different types of data related to the machines is described in the next chapter.

The following chapters summarise the contributions produced during the project. The first contribution of the thesis focuses on building a link between the various types of data. In particular, it takes into consideration the process and output data, and it focuses on the identification of relevant process variables that mostly affect the quality. It proposes a variable selection algorithm based on the Random Forest model to address the case of input variables highly correlated. The second contribution introduces a hybrid approach that combines correlation analysis using observational data and machine learning techniques with designed experiments to establish causality. This approach is particularly useful when investigating unknown phenomena with a large input space, enabling insights and correlations to be gathered from observational data before conducting experimental designs. Next, a novel methodology for analyzing the printer's process data that are organized in a three-way array of

data with a multi-group structure is presented. This methodology is applied to the problem of process modelling, where batches are structured according to multiple groups. By extending the PARAFAC model, the proposed approach accounts for the grouping structure of three-way data sets, enabling the estimation of a model representative of all the groups simultaneously. Following that, a model for the supervised analysis of multi-group three-way data and multi-group output data is described. Finally, some of the collaborative projects conducted with the research group during the Ph.D. project are presented.

Overall, this thesis addresses the challenges and limitations of additive manufacturing processes, contributes to the understanding and optimization of 3D printing, and provides innovative methodologies for data analysis in manufacturing applications.

Preface

The present thesis has been prepared at the Department of Applied Mathematics and Computer Science, Section for Statistics and Data Analysis at the Technical University of Denmark (DTU). It represents one of the requirements for acquiring a Doctor of Philosophy (Ph.D.) degree in Applied Statistics.

The Ph.D. project was funded by the Manufacturing Academy of Denmark (MADE) and has been completed under the guidance of Professor Murat Kulahci (main supervisor), Professor Jesper Henri Hattel and Senior Researcher David Bue Pedersen (co-supervisors). The external research stay has been conducted at KU Leuven, Department of Biosystems, MeBioS division in Belgium, under the supervision of the research manager Bart De Ketelaere. The external research stay has also been funded by the Otto Mønsteds Foundation.

The Ph.D. thesis deals with data analysis methods intended for the new emerging additive manufacturing technologies. All the included work has been carried out during the Ph.D. study period September 2020 - September 2023.

Kongens Lyngby, 31st August 2023

Marta Rotari

Marta Rotari

Acknowledgements

This thesis concludes the three-year Ph.D. program. First and foremost, I am glad and lucky to have been given the chance to study, learn, and communicate science. I always believed that science gives you a distinct delight and joy.

I want to express my gratitude to Professor Murat Kulahci for giving me this amazing opportunity to collaborate on a very interesting project that has always intrigued me and stimulated my curiosity. I would like to thank him for being a good supervisor who advised and taught me with instructive and valuable comments that helped me grow as a person and as a researcher. I would also like to thank my co-supervisors and the whole industrial partner team for the comments, feedback and recommendations they gave me throughout the project. Also, a special thank you to the entire research team, Hao-Ping, Shuo and Kenneth, with whom I spent a lot of time working on various projects and attending several workshops and industrial trips. I shared and learned a lot from and with them. Thank you for all the laughter, conversations, lunches and dinners over the last three years.

Many thanks also go to the entire Section of Statistics and Data Analysis. It is a really nice team to be part of! I am grateful for all the lunches, coffees and activities we shared these years.

During the spring of 2022, I spent the period of my external stay at the University of KU Leuven under the supervision of Bart De Ketelaere. I would like to thank him very much for giving me this fantastic opportunity to visit his Department and meet many fantastic people who made me feel at home from the first moment. Thank you for the fantastic time I spent there, of which I have wonderful memories. A special thank you goes to Valeria, with whom I worked at KU, an excellent researcher and a friend.

A very special thank you to my entire family, especially my mom, for supporting me during these years and helping me through the ups and downs; this thesis is also thanks to you. I would not be who I am, and I would not be where I am without Idriss's love and affection. Thanks for the courage and determination you give me every day. Thank you for celebrating with me all the small and big successes and for being so supportive even during difficult times.

List of publications

The following papers are included in the thesis:

- Paper A: *"Variable selection wrapper in presence of correlated input variables for random forest models"*, M. Rotari and M. Kulahci, Quality and Reliability Engineering International (2023).
- Paper B: *"From correlation to causality"*, M. Rotari and M. Kulahci, to be submitted at Quality Engineering.
- Paper C: *"An extension of PARAFAC to analyze multi-group three-way data"*, M. Rotari, V. Fonseca Diaz, B. De Ketelaere and M. Kulahci, submitted to Chemometrics and Intelligent Laboratory Systems.
- Draft Paper D: *"Multi-way PLS for the analysis of multi-group three-way data"*, M. Rotari and M. Kulahci.
- Paper E: *"In-process monitoring of selective thermoplastic electrophotographic process by laser profiling system and digital fingerprint"*, S. Shan, H.-P. Yeh, M. Rotari, K. Ælkær Meinert, J. H. Hattel, D. B. Pedersen, M. Kulahci, H. N. Hansen, Y. Zhang, and M. Calaon, 23rd international conference of the european society for precision engineering and nanotechnology.
- Paper F: *"Thermo-mechanical model for a selective thermoplastic electrophotographic process for dimensional defects"*, H.-P. Yeh, M. Rotari, S. Shan, K. Ælkær Meinert, J. H. Hattel, M. Kulahci, D. B. Pedersen, and M. Calaon, 23rd international conference of the european society for precision engineering and nanotechnology.

The work carried out in connection with the Ph.D. study has been presented at different conferences (ordered chronologically):

- M. Rotari and M. Kulahci, "Deciphering Random Forest models through conditional variable importance", The 21th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS), Online Conference 2021. (Abstract and presentation)

- M. Rotari and M. Kulahci, "Analysis of multi-group data in a three-way structure", The 22th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS), Trondheim, Norway, June 2022. (Abstract and presentation)
- M. Rotari, H.Yeh, S.Shan, K. Meinert "Four Ph.D. students – one goal: To understand and improve 3D-print", MADE Innovation Conference: Join forces with others and come up with ideas and solutions, Copenhagen, Denmark, August 2022. (Presentation)
- M. Rotari and M. Kulahci, "Analysis of a three-way array that presents a Multi-group structure", Danish Society of Chemometrics (DSK), Middelfart, Denmark, November 2022. (Presentation)
- M. Rotari and M. Kulahci, "Production Analytics for a novel additive manufacturing system", Young Researchers Workshop, Roskilde, Denmark, November 2022. (Presentation)
- M. Rotari and J. Hattel, "Tomorrow's 3D printing and AM solutions", The annual MADE event 2023, Copenhagen, Denmark, May 2023. (Presentation)
- H.Yeh, M. Rotari, S.Shan, K. Meinert and M. Kulahci, "Multiphysics Modelling for Polymer-based Additive Manufacturing Technologies", The Spring Meeting Conference of the European Network for Business and Industrial Statistics (ENBIS), Copenaghen, Denmark, May 2023. (Abstract and presentation)
- M. Rotari, H.Yeh, S.Shan, 23rd International Conference & Exhibition, Copenaghen, Denmark, June 2023. (Posters)
- M. Rotari, M. Kulahci, H.Yeh, S.Shan, 23rd International Conference & Exhibition, Copenaghen, Denmark, June 2023. (Posters)

Contents

CHAPTER 1

Introduction

Throughout its history, the manufacturing industry has experienced several significant technological advancements, resulting in a fundamental restructuring of its operations. In the 19th century, the industry emerged by employing the power of water and steam. Subsequently, it transformed with the advent of electricity and then again with the introduction of computers, which marked yet another significant milestone. The manufacturing sector is now enduring a rapid transition towards digitalization, also known as the fourth industrial revolution or Industry 4.0. Industry 4.0 seeks to transform conventional factories into smart factories equipped with sensors and autonomous systems, which will enable automation and data-driven operations. The Internet of Things (IoT) for data manipulation and communication, augmented reality (AR) and cloud computation are driving this revolution. Innovations such as 3D printing and advanced data analytics are transforming the manufacturing landscape, setting the way for greater industry-wide efficiency, productivity and connectivity.

The manufacturing industries of plastic objects currently revolve around the predominant production through injection moulding. The combination of mass production capabilities and high precision offered by injection moulding has made it a cornerstone of the manufacturing industry. Its versatility allows for the production of a wide range of plastic objects in various sectors, such as automotive, electronics, consumer goods and medical devices. The production process for injection moulding starts with the creation of the mould insert. Next, the plastic material is heated and the molten plastic is injected into the mould cavity through a specially designed nozzle. Once the plastic has cooled and solidified, the mould is opened, and the newly formed plastic object is ejected. Despite this process apparent simplicity and speed, which enables the rapid production of large quantities, it is essential to recognize certain limitations associated with injection moulding. Notably, mould production necessitates a significant investment of time and numerous iterations to achieve optimal quality results of the final products. Moreover, the requirement to fabricate individual simple parts and subsequently assemble them to obtain a more intricate final product represents another constraint. These limitations can pose challenges to manufacturing companies requiring intricate and highly detailed components or fast prototyping, demanding alternative production methods to achieve the desired results.

Companies are driven towards continuous innovation in today's rapidly and dynamic evolving landscape. Innovation is often fueled by the demands of the 21st

century and the needs of contemporary consumers, encompassing diverse factors such as the need for material differentiation, customized products, ever-changing and progressively intricate colour schemes and complex shapes. Companies are driven to constantly explore new colours, textures and materials combinations that captivate consumer interest and align with the prevailing trends. Sustainability has simultaneously become an urgent concern for both enterprises and society. Companies actively seek ways to minimize waste, explore novel materials and new and more efficient production processes to ensure a greener future. Numerous investments are made in research and development to discover innovative ways of optimizing resource utilization, designing recyclable and biodegradable materials, and adopting circular economy principles.

The convergence of factors such as material differentiation, customized products and diverse colour schemes, alongside the increasing importance of sustainability, drives industries to allocate resources towards research and innovation. In response, companies actively seek new production processes that can address these challenges and offer greater flexibility and efficiency. One notable response to this demand is Additive Manufacturing (AM), also known as 3D printing. The industries explore the potential of additive manufacturing as a transformative force in the industry. This technology enables the creation of complex shapes and customized products with reduced material waste, aligning with sustainability principles. Furthermore, the digitization of manufacturing processes facilitates automation, data-driven decision-making and seamless integration of systems, providing companies with the tools to drive innovation and remain competitive in a rapidly evolving landscape. By combining research, innovation and adopting advanced production processes like additive manufacturing, companies can enhance synergies that lead to sustainable growth, efficiency, productivity and flexibility.

The manufacturing industry in Denmark is witnessing a remarkable shift towards digitization and the adoption of new production methodologies. The collaborative efforts of Danish companies through the consortium Manufacturing Academy of Denmark (MADE) for Industry 4.0 exemplify the proactive approach taken by the industry to leverage digital technologies, drive innovation and secure their position as global leaders. Through collective knowledge sharing, research collaborations and talent development, MADE aims to drive the Danish manufacturing sector into the digital era, fostering sustainable growth, competitiveness, and prosperity.

1.1 Project environment

The pursuit of innovation necessitates continuous adaptation and improvement. Companies are driven to explore cutting-edge technologies, processes and materials that enable them to differentiate their offerings. In Denmark, numerous companies have taken a proactive approach by forming a collaborative consortium known as MADE (Manufacturing Academy of Denmark) for Industry 4.0, with a primary focus on driv-

ing digitization and fostering innovation within the manufacturing sector. The advent of Industry 4.0 represents a paradigm shift in the manufacturing landscape, characterized by integration of digital technologies, automation and data-driven decision-making.

The consortium's primary objective is leveraging digitization and innovation to maintain the country's global industry competitiveness. MADE focuses on accelerating the adoption of Industry 4.0 principles in Danish manufacturing. The consortium members collaborate on research and development, sharing best practices and collaborating on cutting-edge research and development initiatives. The strategic initiatives of MADE encompass a wide range of activities to facilitate the digital transformation of the industry. This includes implementing smart manufacturing technologies like IoT, AI, big data analytics and cloud computing. These technologies optimize production processes, enable real-time data-driven decision-making, and enhance operational efficiency. The consortium also promotes collaboration between industry and academia, fostering knowledge exchange, talent development and a skilled workforce capable of driving digital transformation. MADE serves as a platform for disseminating knowledge and showcasing successful digital transformation case studies, inspiring other companies to embark on their digitization journeys.

Figure 1.1. MADE Fast structure, `https://www.made.dk/en/made-fast/`.

The MADE consortium was founded in 2014 in conjunction with the Spir Project, which ran from 2014 to 2019. Building upon the success of the Spir Project, the consortium embarked on another significant initiative called MADE Digital, which lasted between 2017 and 2020. MADE has started its most recent initiative by introducing MADE FAST (Flexible, Agile, and Sustainable Production enabled by Talented Employees). This new initiative represents an industrial-led research, innovation and education partnership to promote the development of the next generation of advanced manufacturing capabilities in Denmark. MADE FAST encompasses five

distinct workstreams, each dedicated to specific aspects of research and development, as illustrated in Figure 1.1. These workstreams serve as focused areas of exploration and collaboration, fostering advancements in manufacturing technologies, processes and human talent.

This PhD project specifically falls under Workstream 4, which is focused on achieving sustainable upscaling through the digitalization of manufacturing processes. The main objective of Workstream 4 is to optimize production processes and enhance product quality by leveraging digital tools and data-driven models. These technological advancements aim to reduce the time required to ramp up new production and ensure a "first-time-right" approach to manufacturing new products. Under the Workstream 4 umbrella, a large initiative comprising four interconnected PhD programmes has been initiated. The initiative's primary objective is to investigate two novel additive manufacturing techniques: the Selective Laser Sintering (SLS) and the Selective Thermoplastic Electrophotographic Process (STEP). The project's overall goal is to better understand these two novel production processes and promote their digitization and optimization so that they can be scaled up to a production line. Figure 1.2 illustrates the collaborative nature of these projects, in which the numerous PhD projects work together to accomplish common goals.

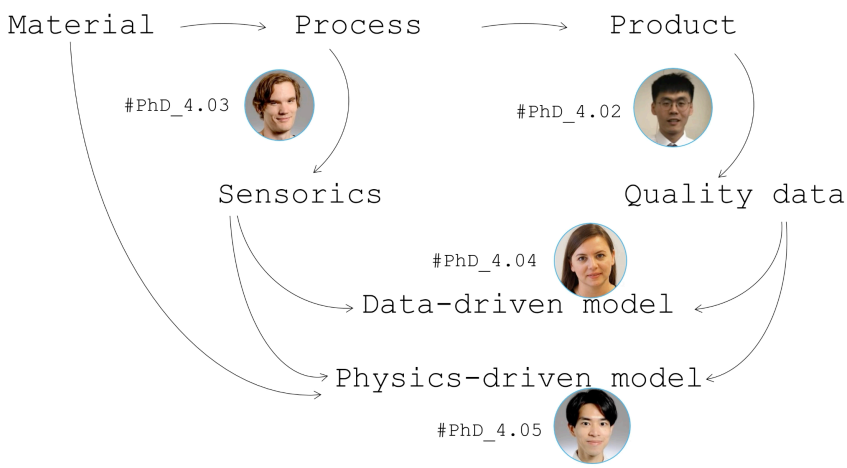


Figure 1.2. MADE Workstream 4 project on Additive Manufacturing.

Project 4.03 mainly focuses on sensoring the machines. The sensors, strategically located, collect and analyze experimental data fed into the data-driven and physical-driven models. A second project is Project 4.02, which focuses on digital fingerprint analysis aimed at enriching the understanding of product qualities. The data collected by the sensors, the input parameters and the quality of the final products are

analyzed in the current Ph.D. Project 4.04 develops data-driven models aimed at process understanding and optimizations. Material and quality aspects, with process information, are used in the Multi-physics model built in Project 4.05, aimed at process digitalization and optimization from the physics point of view. The final aim of the combination of all projects is to use the data and physically-driven model results to be fed back into the process for optimization.

1.2 Project objectives

In the context of technological advancements and innovative production processes, additive manufacturing has emerged as a noteworthy solution. Additive manufacturing, also known as 3D printing, is a process of making a three-dimensional object from a digital model. Two additive manufacturing processes were taken into consideration: the Selective Laser Sintering (SLS) and the Selective Thermoplastic Electrophotographic Process (STEP). Both technologies are described more thoroughly in Chapter 2. Both technologies are described more thoroughly in Chapter 2.3. SLS and STEP printers have huge potential to revolutionize production processes, however, their current status requires extensive research and development efforts to upscale them for industrial applications.

These technologies are relatively nascent in the global market and limited knowledge exists regarding their full potential. To achieve the industrial-scale implementation of 3D printing, it is imperative to address several key challenges. Firstly, the current understanding of the 3D printing process is limited, with scarce literature available. Therefore, a primary focus is placed on expanding the knowledge base of the process itself and comprehending its underlying mechanisms. Secondly, the optimization of the entire 3D printing process is crucial to ensure reproducibility and achieve consistently high-quality output. Lastly, one of the significant challenges hindering the integration of 3D printing into large-scale production is its inherent instability. The process often exhibits fluctuations and inconsistencies, resulting in unpredictable outcomes.

Multiple sources of data can be identified within the additive manufacturing (AM) process. The **input data** includes machine settings; the **process data** represents the data acquired during the printing process through multiple sensors allocated through the production chain. Finally, the **output data** includes the quality of the final products. These data categories are described more thoroughly in Chapter 2.5.

The primary objective of this project is to enhance the understanding and comprehension of various technologies using data-driven models. To achieve this goal, the project employs descriptive machine learning, statistical analysis, and chemometrics models. By utilizing these methodologies, the project seeks to increase knowledge and awareness of the technologies. This entails establishing potential connections and meaningful linkages among the diverse data sets involved in the processes (see Figure 1.3). The analysis aims to identify key variables that significantly influence

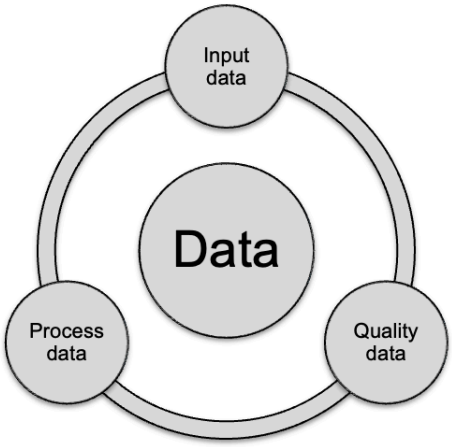


Figure 1.3. Data overview.

the production process and their impact on the quality of the final products. Furthermore, the project aims to develop new methodologies to analyze and explore the various structures of the data collected during the printing process.

Another crucial objective of the project is process optimization, which involves determining the optimal settings of machine variables to ensure to achieve the desired outcomes. In essence, the goal is to identify the input parameters that exert the most significant influence on product quality. By establishing causal relationships between the input and output spaces, the project aims to pinpoint the variables and their optimal levels necessary to ensure the desired quality consistently.

Through a comprehensive data-driven approach, this project strives to enhance the knowledge and comprehension of the targeted technologies. By employing various analytical models, it seeks to gain valuable insights into the processes, optimize them for improved outcomes, and ultimately contribute to advancing the efficiency and effectiveness of these technologies in real-world applications. Furthermore, the project entails interacting with other members of the group project, often engaging in joint initiatives and identifying areas for collaboration based on shared interests and the need for specialized knowledge. The team integrates the discoveries of one another into their models, such as inserting additional sensors in need of more data or comparing and improving mutual models and findings by validating results and work.

1.3 Contributions

At the beginning of the project, the first focus was placed on establishing a connection between various types of data. In this instance, a machine learning model was

developed with a twofold objective. The primary aim was to analyze the relationship between process variables and the resulting output data, specifically identifying the few process variables that exert the most significant influence on product quality. In this regard, instead of merely utilizing a ”black-box” model for response prediction, our approach emphasized the identification of relevant input variables throughout describing the contributions of these input variables in the form of variable importance. An additional challenge encountered was the presence of highly correlated process variables, which could introduce bias in calculating their importance scores, as highlighted in the preliminary literature review. To address these objectives, a tree-based machine learning model, Random Forest, was employed along with a special variable importance measure that takes into consideration the correlation among data. The second objective was to identify and retain only the most meaningful variables, omitting those with minimal or negligible impact on the product quality. In other words, based on the variable importance scores ranking, the goal was to find a cutoff or a threshold level of variables to retain. To achieve these objectives, an extension of the wrapper algorithm, Boruta, was developed to handle the case of highly correlated variables. This extended model efficiently selected the most relevant variables while discarding those with negligible impact. The findings and outcomes of this study are presented in Paper A.

The initial model served as a first step to investigate the causal relationship between input and output data. With the emergence of advanced technologies, production processes have undergone significant customization, resulting in increased complexity within production systems. This complexity arises from the integration of numerous process parameters, thereby substantially expanding the input space. Consequently, identifying the specific input variables in such a vast space that has a causal relationship with the output variable becomes a challenging task. In this domain, an approach that addresses the problem of transitioning from correlation analysis to assessing causality within a vast input space is presented. Analyzing causal relationships becomes increasingly complex as the number of variables grows. Merely identifying correlations between variables is insufficient to establish cause-and-effect relationships definitively. The proposed approach offers a systematic method to move beyond correlation and delve into the realm of causality. Combining the use of machine learning techniques and subsequent controlled experiments, this approach aims to identify the specific input variables that truly drive the changes in the output variable, thus establishing the causal links. It provides insights into the key factors that significantly impact product quality, leading to more targeted interventions and process optimizations. The results related to this approach are presented in paper B.

The primary objective of new technologies is to facilitate large-scale production by enabling the creation of numerous items in a single production cycle. The production process itself unfolds in a structured manner, with products being manufactured in successive rows. During the production phase, the process data are collected from an array of sensors strategically positioned throughout the entire production process. These sensors capture valuable information at each stage of the production cycle. The association between the observations in the process data and the rows of products

creates a multi-group structure in the data. An objective of the project was the development of new methods for analysing the data collected by these technologies. We developed a new methodology that addresses this multi-group structured three-way array effectively. This model follows within the Factor analysis, extending the existing Parallel Factor Analysis (PARAFAC) model. By applying this model, we enriched our understanding and analysis of the process data. The results corresponding to this new methodology are presented in paper C.

A successive methodology development involved the extension of the multi-group three-way method to accommodate the output data, thereby enabling a transition from unsupervised to supervised analysis. This extension takes into consideration the multi-group structures of the process data, which allowed us to consider different groups within the dataset and their respective relationships with the quality matrix. The basis for this methodology relates to the multivariate Partial Least Squares (PLS) model. The primary objective of this algorithm was to investigate the relationship between the multi-group three-way array representing process data and the quality matrix. This development aims to improve our comprehension of the complex relationship between process data and the associated quality matrix, thus establishing the way for more informed and insightful analyses. This new methodology is presented in the draft Paper D.

The PhD project entails a collaborative effort with PhD students involved in the same MADE research project. As a team, we have collectively dedicated our efforts to investigate two distinct projects centred around various aspects within the domain of 3D machines. The first project within this collaborative framework focuses on the monitoring of layer-by-layer production utilizing laser-based techniques. The laser serves the purpose of tracking the precise deposition of layers onto the main component. Subsequently, an in-depth analysis of the images obtained through the laser enables us to analyze the production process over time. This project also opens up the possibility of monitoring production in real-time. The findings and outcomes derived from this project are presented in Paper E. The second project in joining collaboration focuses on investigating and addressing the dimensional defects of the final products produced by the 3D machine. Through this collaboration, we have successfully developed models, conducted research activities and accumulated valuable insights concerning the identification and characterization of the dimensional defects. The analysis and results derived from this project have been presented in paper F.

1.4 Outline of the thesis

This thesis project aimed to improve the understanding and quality of 3D printing processes through advanced data analysis and the development of new methodologies. The research was conducted in the context of the MADE research project, focusing on various aspects within the domain of 3D machines. The main objectives and outcomes of the thesis are organized as follows:

- A brief introduction to additive manufacturing is given in chapter 2. The chapter includes a brief description of the new 3D printing technologies, their benefits, limitations and data organization.
- In chapter 3 the methodology developed for variable selection in case of high correlation among data is presented and includes Paper A.
- An approach to address the challenging task of transitioning from correlation analysis to assessing causality within a vast input space is presented in chapter 4. The chapter also includes Paper B.
- A new methodology to analyze three-way arrays with a multi-group structure is presented along with Paper C in chapter 5.
- Multi-way N-PLS model for three-way data with a multi-group structure is presented in chapter 6.
- Chapter 7 contains the two papers developed together with the research group, specifically Paper E and F, and other data analyses regarding the new technologies.
- Finally, the conclusion on the main findings and contributions of the thesis, along with future works and final remarks on the research are presented in chapter 8.

CHAPTER 2

Industrial environment

2.1 Manufacturing processes

In the modern age, we are surrounded by objects, many of which are made of plastic. This raises the question: how are these objects manufactured? The process of injection moulding primarily dominates the production of plastic objects. This manufacturing technique has emerged as the dominant method for creating plastic products, enabling mass production with remarkable efficiency and versatility. The origins of injection moulding can be traced back to the late 19th century when early attempts were made to mould celluloid, an early form of plastic. However, it was not until the mid−20th century that significant materials science and machinery advancements propelled injection moulding into mainstream manufacturing. With the introduction of thermoplastics such as polypropylene and polyethene, the technique became more refined, resulting in its widespread adoption across industries.

In injection moulding, a meticulously coordinated sequence of steps is employed to fabricate plastic objects. Figure 2.1 is a visual representation of the injection moulding procedure. Central to this technique is the utilization of mould inserts, which play a crucial role in the overall process. It involves the meticulous design and fabrication of a custom-made mould component that defines the shape and details of the final plastic object, ensuring precision and accuracy in its reproduction during production. The process commences with carefully selecting and preparing raw materials, typically in the form of granules or pellets. These materials are then melted within a specialized injection moulding machine, where they are subjected to high pressure and injected into a precisely designed mould cavity. Following this injection, the molten plastic rapidly cools and solidifies within the mould, forming the desired object. Finally, the mould is opened, and the newly produced item is ejected, ready for further finishing or assembly processes.

The dominance of injection moulding products can be attributed to its numerous advantages and benefits. The mould's versatility allows for the production of items ranging from small components to large-scale products. Injection moulding enables high production rates, ensuring the rapid and cost-effective manufacture of plastic items [1]. Moreover, this process offers remarkable precision and repeatability, ensuring consistency in the quality and dimensions of the final products [2]. The use of automated machinery further enhances efficiency, reducing labour-intensive tasks and human error. Lastly, injection moulding permits the utilization of a wide

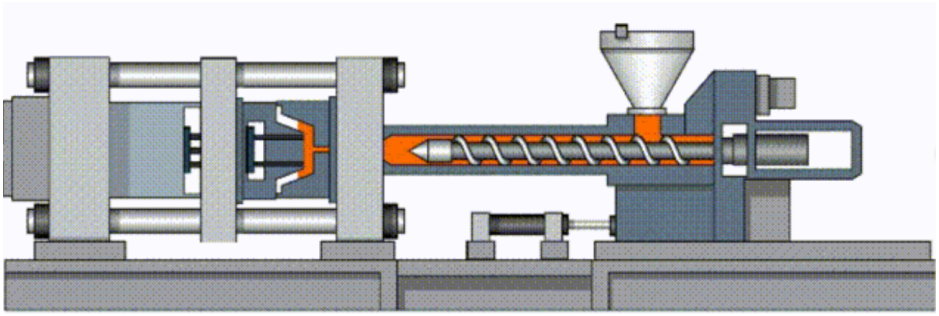


Figure 2.1. Representation of the injection moulding process.

range of thermoplastic materials, each possessing unique properties, such as strength, flexibility and resistance to heat, chemicals or UV radiation [3].

Although injection moulding is renowned for its apparent simplicity and rapid production capabilities, it is crucial to acknowledge the inherent limitations associated with this process. One such limitation is the considerable investment of time and iterative refinement required for mould production to attain optimal results. Designing and fabricating a mould necessitates careful consideration of various factors, including geometry, material selection and cooling system design. Iterative adjustments and testing are often required to fine-tune the mould design, which can prolong the production timeline. The production of the mould insert requires significant investments in terms of time, costs and iterations. This fact determines the use of the mould insert for mass production, limiting the possibility of producing a mould for small productions or prototypes. Additionally, the resulting parts from injection moulding tend to be simple in nature, lacking intricate details or complex shapes. Intricate components may require additional assembly steps or post-processing to achieve the desired level of detail. This assembly process adds complexity, cost and time to the overall manufacturing process. These limitations can pose challenges for manufacturing companies that require highly detailed and intricately designed components or seek fast prototyping. In such cases, alternative production methods, such as 3D printing, may be considered to achieve the desired results more efficiently.

Additive Manufacturing (AM) has emerged as a noteworthy solution in the realm of technological advancements and innovative production processes. Also referred to as 3D printing, this cutting-edge technology presents novel production prospects and broadens the scope of possibilities within the manufacturing domain. This new manufacturing process represents a departure from traditional manufacturing methods, such as machining or moulding. The versatility and potential of 3D printing are of interest to numerous industries, ranging from aerospace and automotive to healthcare and consumer goods.

The AM process begins with a digital representation of the three-dimensional object. The digital model is then divided into distinct layers, a process known as slicing, which forms the foundation for the subsequent stages of production. This

innovative technique allows for gradually adding material layer by layer, resulting in the fabrication of three-dimensional objects. AM complements traditional methods by offering unparalleled flexibility, customization, and on-demand production. While not replacing conventional methods like injection moulding, additive manufacturing offers a complementary approach that enriches the manufacturing landscape. As this technology advances, it can revolutionize numerous industries, foster innovation and redefine production in the modern world.

2.2 Why additive manufacturing?

The advent of AM can transform and redefine how we conceive, create, and produce objects in the modern world. In this section are described in more detail some of the benefits AM offers and the novel capabilities and frontiers it enables.

Numerous materials One of the most remarkable aspects of additive manufacturing is the wide range of materials that can be utilized for printing. Traditional manufacturing processes often require specific materials and elaborate tooling setups, constraining the flexibility and adaptability of production. Conversely, 3D printing enables the use of various materials, including plastics, metals, ceramics, and even composites, enabling the creation of multi-material products. This capability to work with diverse materials broadens the design possibilities and offers opportunities for functional optimization, lightweight and enhanced performance.

Intricate geometries Additive manufacturing is a breakthrough for intricate geometries. The layer-by-layer approach allows for intricate internal structures previously unattainable through traditional methods. Complex geometries, organic shapes and interlocking parts can be realized, enabling new design paradigms and innovative product concepts. This newfound freedom enables engineers to explore novel solutions, pushing the boundaries of what is conceivable and achievable in manufacturing. Moreover, additive manufacturing offers the capability to create tools, enabling rapid exploration of various tool designs. This agility empowers manufacturers to swiftly prototype and test new tooling concepts, significantly reducing the time required for tool development and facilitating expedited evaluation of their performance and functionality.

Reduced time from prototype to production Beyond the design possibilities, additive manufacturing brings significant advantages to the production process itself. One of the notable benefits is the reduced lead time from design to final product. Traditional manufacturing often involves lengthy and costly tooling processes, necessitating substantial investments of time and resources before production can commence. In contrast, 3D printing eliminates or significantly minimizes the need for tooling, enabling rapid prototyping and shortening the time to market for new products. This

accelerated development cycle fosters innovation and agility within industries, allowing quicker responses to market demands and customer requirements. It provides an avenue for cost-effective small-batch manufacturing and customized production that were previously unviable.

Multiple colours Multiple colours can be incorporated into a single printed component through additive manufacturing. The resulting objects can exhibit a wide range of palettes by integrating colour transitions, gradients, and shades. These advances in AM technology have revolutionized colour customization. With conventional manufacturing techniques, attaining such elaborate colour effects frequently requires complicated post-processing techniques or distinct assembly steps. However, additive manufacturing offers a more streamlined approach, integrating the colouring procedure directly into the object's fabrication.

It is worth emphasizing that AM should not be perceived as a substitute for conventional manufacturing techniques, such as injection moulding, but rather as a supplementary approach that expands the realm of possibilities. Injection moulding, with its efficiency and high-volume capabilities, continues to play a crucial role in mass production. Nonetheless, additive manufacturing can complement this traditional method by offering flexibility, customization, multi-material and on-demand production. It presents a cost-effective avenue for small-batch manufacturing, tailored and personalized production, and the realization of intricate designs previously deemed unfeasible. AM can represent a step in the evolution of production.

2.3 Two additive manufacturing processes: SLS and STEP

In AM, two prominent technologies have emerged as valuable candidates for scaling up to industrial production lines: Selective Laser Sintering (SLS) and Selective Thermoplastic Electrophotographic Process (STEP). This section contains a concise explanation of these technologies.

2.3.1 The Selective Laser Sintering

SLS is an additive manufacturing process that uses a high-powered laser to selectively melt and fuse together particles of powdered material, Figure 2.2. The SLS process can be described through the following several key steps:

1. **Pre-processing**

    - **Design**: The first step in the SLS process is to create a 3D CAD model of the object to be manufactured. This model is then sliced into thin layers,

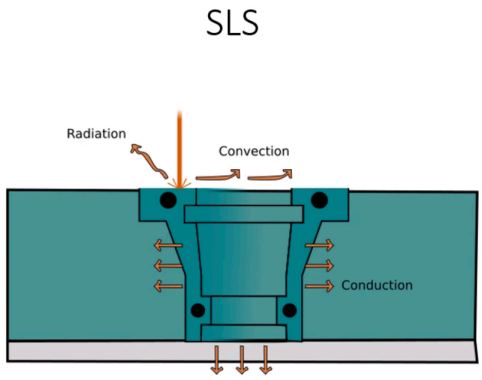


Figure 2.2. Selective Laser Sintering.

typically between 0.013 and 0.016 mm thick. This prepares the design for the additive manufacturing process.

- **Material selection**: A powdered material is selected based on the desired properties of the final product. SLS is compatible with a wide range of materials, including plastics, metals and ceramics.

2. **Printing process**

   - **Powder bed**: A thin layer of the powdered material is spread evenly over a build platform. This layer is typically between 0.1 and 0.3 mm thick.
   - **Laser scanning**: A high-powered laser scans the surface of the powdered material, selectively melting and fusing together particles in the shape of the first layer of the CAD design. The laser is controlled by a computer, which uses the CAD data to precisely position the laser and control the intensity and duration of the laser pulses.
   - **Layer-by-layer**: Once the first layer has been sintered, the build platform is lowered by the thickness of one layer and the process is repeated to add successive layers until the final object is complete. Each layer is scanned and sintered in the same way as the first layer, with the laser selectively melting and fusing together particles in the shape of the next layer of the CAD design.
   - **Cooling and removal**: Once the object is complete, it is allowed to cool and harden.

3. **Post-processing** The excess powdered material is removed, usually using air jets or special washing, leaving behind the final product. Any remaining powder

can be reused in the next SLS build, making the process more efficient and cost-effective.

SLS can be used to create complex geometries and intricate details that are difficult or impossible to produce using traditional manufacturing methods, allowing for flexible production. It can produce parts quickly and efficiently, with minimal setup time. It is cost-effective as it eliminates the need for expensive moulds or tooling, making it a more cost-effective manufacturing option for small production runs or custom parts. Furthermore, SLS is compatible with a wide range of materials, including plastics, metals, and ceramics, making it a versatile manufacturing process for various applications.

2.3.2 The Selective Thermoplastic Electrophotographic Process

The Selective Thermoplastic Electrophotographic Process (STEP) is a type of additive manufacturing technology that uses electrostatic forces to deposit and fuse layers of thermoplastic powder, producing completely dense multi-material. It is a brand-new polymer-based additive manufacturing method introduced by Evolve Additive Solutions, Inc, Figure 2.3. The STEP technology is mainly composed of two fundamental modules, electrophotographic engine and transfusion module as shown in Figure 2.4. The printing process can be summarized in the following key steps:

Figure 2.3. STEP introduced by EVOLVE.

1. **Pre-processing**

    - **Design**: The STEP process starts with the creation of a 3D CAD model of the object to be manufactured. The model is then sliced into several nearly 2D layers. The thickness of the layer is around 0.013 mm to 0.02 mm.

- **Material selection**: A thermoplastic powder is selected based on the desired properties of the final product. STEP is compatible with a range of thermoplastic materials, including nylon, polycarbonate and polypropylene.

2. **Printing process**

   - **Electrostatic printing**: The belt, responsible for transporting the layers, receives distinct charges from the electrophotographic module based on the CAD data, ensuring precise positioning of the individual charges. Depending on the specific charges received, the material is selectively deposited on the belt, serving as either support or part material.
   - **Heating**: The belt, with the layer on top, travels beneath a heat-emitting lamp, which increases the material's temperature in preparation for fusion. Once the layer has been appropriately heated, it is deposited on the building plate.
   - **Fusion and cooling**: The building plate moves horizontally in three positions: heating, depositing and cooling. In the heating position, the building plate and the previously printed layers undergo heating up toglass transition temperature. As the building plate shifts to the transferring position, the subsequent layer on the belt is melted and fused, transferring onto the building plate. Following the transfer, the building plate proceeds to the cooling position, gradually cooling down. Throughout the printing process, the building plate moves back and forth to facilitate stacking each successive layer upon it.
   - **Layer-by-layer**: Once the first layer has been fused, the process is repeated to add successive layers until the final object is complete. Each layer is deposited and fused in the same way as the first layer, with the electrophotographic module depositing materials on the image of the cross-section and the transfusion module melting the powder particles together.

3. **Post-processing**

   - **Cooling**: Once the object is complete and all the layers are fused and pressed together, the building plate, together with the newly formed 3D bulk allowed to cool and harden.
   - **Removal**: The support material is then dissolved in the chemical wash, while the part material remains unaffected and intact, leaving behind the final product.

Figure 2.4 provides a visual representation of the STEP process. This technology offers several benefits over traditional manufacturing methods. STEP is capable of producing parts with high accuracy and resolution, making it suitable for creating parts with intricate details and complex geometries. It can produce parts quickly

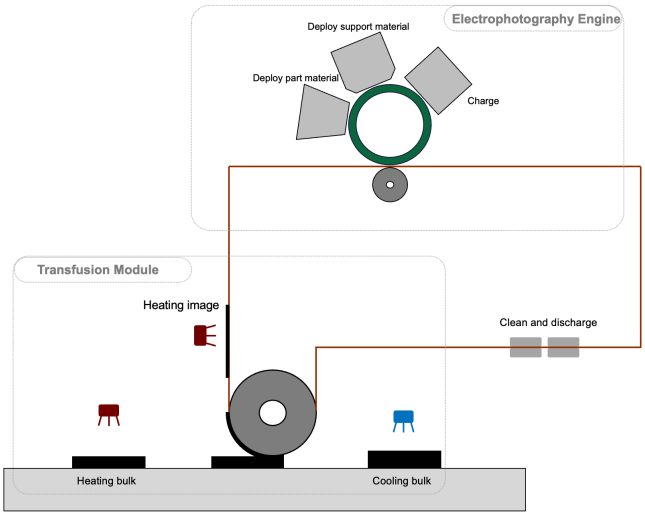


Figure 2.4. Selective Thermoplastic Electrophotographic Process.

and efficiently, with minimal setup time. It also offers several material options and is compatible with a range of thermoplastic materials, allowing for the creation of parts with a variety of properties. Finally, STEP produces minimal waste, as any excess powder can be reused in the next build.

2.4 Limitations and defects and what to improve

SLS and STEP additive manufacturing technologies have emerged as notable solutions in the sphere of technological advancements and innovative production processes. These two prominent printers have the potential to revolutionize production methods. However, substantial research and development efforts are required to scale them up for industrial applications. These technologies are still in their early stages on the global market, and little is known about their maximum potential. To achieve widespread adoption of 3D printing on an industrial scale, it is necessary to address several key challenges.

Firstly, the current understanding of the 3D printing process remains limited, with scarce literature available. Therefore, it is necessary to expand the knowledge base of the process itself, deepen our understanding of its underlying mechanisms and unveil its full range of capabilities.

Additionally, the optimization of the entire 3D printing process is crucial to ensure reproducibility and achieve consistently high-quality output. This involves fine-tuning various machine parameters, analyzing the process data and closely studying the quality of the ultimate products. Developing increased knowledge and models for

machine analysis will aid in minimizing errors, enhancing efficiency and maximizing the overall quality of the printed objects.

Lastly, a significant challenge hindering the integration of 3D printing into large-scale production is its inherent instability. The process often exhibits fluctuations and inconsistencies, resulting in unpredictable outcomes. To address this challenge, further advancements are required in terms of process stability, control systems and real-time monitoring.

The current printing process involves the utilization of building plates capable of fitting multiple products simultaneously, arranged in several rows, as illustrated in Figure 2.5. This example showcases ISO 527 tensile bar specimens, highlighting the general nature of the application. In this specific case, three rows of 17 bar specimens have been produced, although the production capacity can be extended to encompass additional rows.

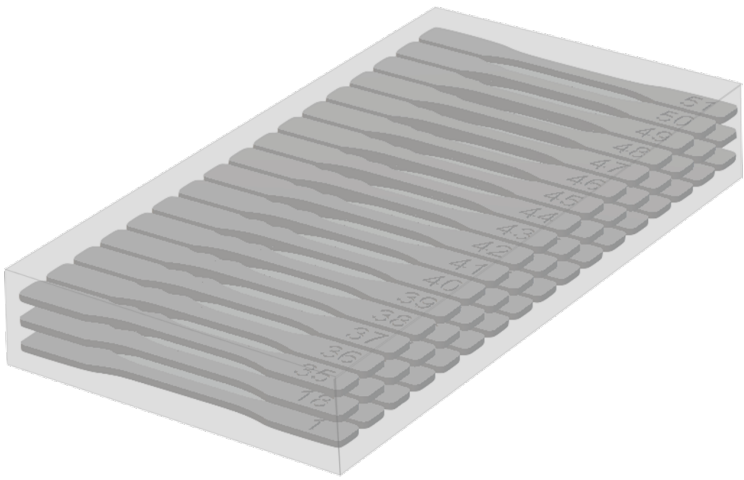


Figure 2.5. STEP final products.

Ensuring a high level of quality across the various products is crucial in such scenarios. Numerous quality aspects are considered, including mechanical characteristics like tensile strength and Young's Modulus. Several defects are analyzed as dimensions conforming to the specifications outlined in the digital design. Other qualities, such as warpage and surface roughness, are also thoroughly examined. It is imperative to achieve and maintain stringent quality standards across all the produced items. These qualities must be consistently guaranteed, exhibiting homogeneity and uniformity within each product within the same row and between all rows. The attainment of these high-quality standards is vital to instil confidence in the reliability and performance of printed products.

2.5 Data Overview

Data organization follows a similar pattern in the STEP and SLS printing machines. While this thesis primarily focuses on the STEP machine, it is important to acknowledge the broader applicability of these data categories to the SLS technology as well. In the context of the STEP process, the data can be categorized into three main classes: input data, process data, and quality data. Each of these data categories is described individually below. Figure 2.6 provides a visual representation of the different categories.

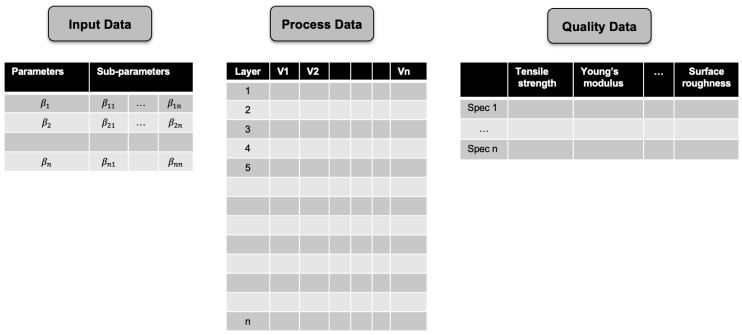


Figure 2.6. Representation of the data of the STEP process.

Input variables. The input data encompasses machine settings or process parameters that are established at the beginning of an order. This data can be represented by a set of multiple parameters, denoted as $\beta_1, \cdots, \beta_n$, wherein each parameter includes a series of sub-parameters $\forall i \in 1, \cdots, n$ can be defined $\beta_{i1}, \cdots, \beta_{im}$ with $m \in \mathbb{N}$.

As an example, we can consider the temperature of the building plate can serve as an initial parameter, β_i, and its sub-parameters can include the lower and upper temperature limits, β_{i1}, β_{i2}. Additionally, subsequent sub-parameters can account for temperature changes as the layer number progresses, $\beta_{i3}, \cdots, \beta_{im}$. For instance, with the accumulation of layers, it may be necessary to decrease the temperature to mitigate any potential issues.

Process variables. Diverse sensors are strategically positioned throughout the production chain to collect data throughout the manufacturing process, comprising a comprehensive set of process data encompassing various production-related information. Specifically, a total of 53 continuous variables are acquired during the production process. The variables are denoted as $V1$ to $V53$ for confidentiality purposes.

For every produced layer, the system records an observation for each variable, thus associating a row within the table depicted in Figure 2.6.

Output or Quality variables. The output data reflects the quality of the manufactured products, which can be of different natures. The products' mechanical qualities include tensile strength and Young's modulus. Other qualities, such as warpage and dimensional defects, were also considered. This includes dimensional measurements to verify if the printed object conforms to the design specifications. Quality data plays a vital role in validating the success of the printing process, ensuring that the printed objects meet the required standards and functional requirements. In general, all the products of the production are tested for their effectiveness and to study how the quality propagates within each row and between rows.

CHAPTER 3
Variable selection

3.1 Introduction

In the initial stages of the PhD journey, significant emphasis was placed on the STEP machine. Our primary objective was to address the challenges associated with the process understanding and establish a means to connect the diverse data set generated by the machine. Specifically, we focused on identifying the process stage that significantly influences the quality. By strategically deploying sensors throughout the machine, we are able to thoroughly investigate the process variables, with a particular emphasis on understanding which variables affect the overall product quality. In this study's context, we considered the process variables as the input variables, while the quality variables were designated as the output variables. In order to address these challenges, descriptive models that can be comprehended are required, rather than solely focusing on predicting responses through "black-box" models. Identifying relevant input variables has emerged as a subject of interest across various domains. Therefore, the growing focus is placed on models that can describe the contributions of the input variables to the model in the form of "variable importance". Certain machine learning techniques, such as random forest (RF), readily provide variable importance measures.

The Random Forest (RF) model is a tree-based ensemble learning algorithm that combines multiple decision trees to make predictions [4]. It is a widely used modelling algorithm for several purposes, including high-dimensional problems with multi-class responses, categorical variables, and imbalanced data and multiple adapted versions have been proposed [5–7]. RF excels in handling complex data sets, as it can effectively capture nonlinear relationships and interactions between variables [8]. Additionally, RF mitigates the risk of overfitting by randomly selecting subsets of features and training each decision tree on different subsets of the data. This diversification leads to robust and reliable predictions [9].

RF provides two variable importance measures, Mean Decrease Impurity (MDI) or Mean Decrease Accuracy (MDA), obtained based on the variable's contribution to the model's performance [4]. Utilizing the variable's importance scores, a rank of the variables can be obtained. This ranking provides an initial indication of the input variables exerting the greatest influence on the output, or, in other words, the most significant variables. While RF provides variable importance measures and a variable ranking, it does not offer guidance on the optimal number of variables to retain as the

most influencing variables. The RF model does not inherently incorporate a threshold or criterion for determining the number of variables to retain, Figure 3.1.

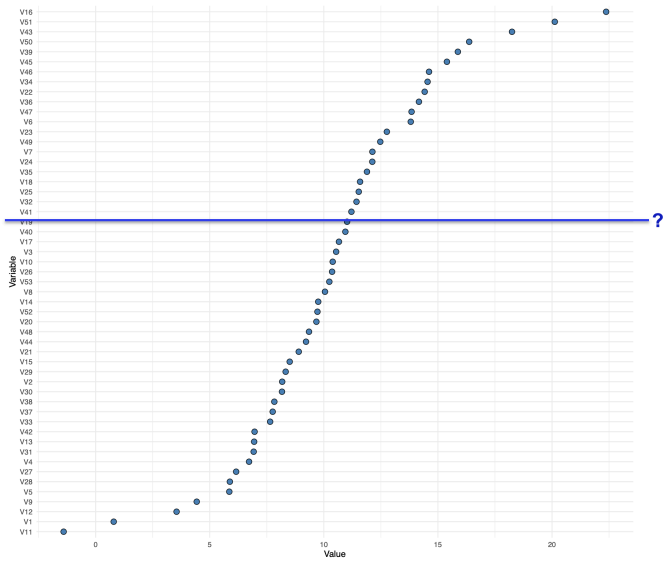


Figure 3.1. Random Forest ranking of the variables based on the MDA variables importance measure.

The absence of a predefined threshold represents a significant challenge in determining the number and specific variables that truly influence the response variable. Finding an appropriate threshold in the ranking of variables obtained from the RF model can be viewed as a variable selection problem. Variable selection is a challenging issue in high-dimensional regression and classification problems. Most of the input variables are irrelevant to predict the output, but their relevance is usually unknown. Typically, there are three goals in variable selection and methodologies might change depending on the aim we pursue. The first aim is to address technical limitations caused by large sets of variables, which slow down algorithms, consume excessive resources and are inefficient. The second aim is to identify a small number of variables that maximize model accuracy, as many machine learning algorithms show reduced accuracy with an excessive number of variables. The third aim is to enhance understanding of the underlying data generation process by identifying and prioritizing all influential variables, particularly when the goal is to gain insight into mechanisms related to the subject of interest rather than solely building predictive models.

One of the challenges encountered when applying the Random Forest (RF) model and subsequently its variable's importance measures is the presence of high correlation among the input variables. Figure 3.2 shows the process data of the STEP

machine, where we can observe numerous groups of variables which show a high correlation among them. This high correlation can impact the conventional RF variable importance measures, leading to an overestimation of the importance scores assigned to the variables [10, 11]. As a result, the ranking of variables based on these measures may not accurately reflect their true significance in influencing the output variable. This phenomenon results in an incorrect ranking, which leads to an incorrect selection of the most relevant variables.

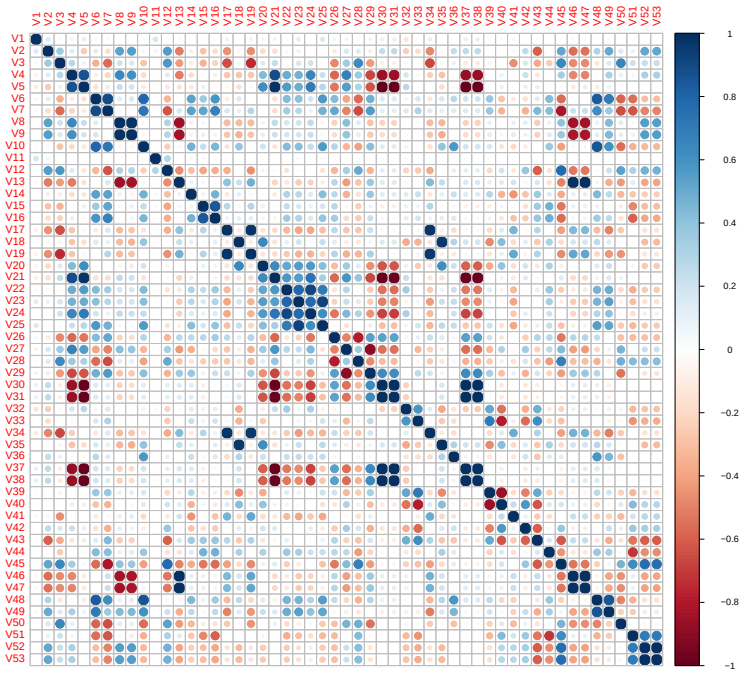


Figure 3.2. Highly correlated process data.

In the literature, there is a noticeable gap in variable selection techniques that adequately address the presence of highly correlated data. This gap has prompted the need for the development of novel methodologies that can effectively handle this challenge. A methodology is presented in the journal article included in the Section 3.3 to address this gap. In situations where there is a high correlation between input variables, the model provides a valuable variable selection technique.

3.2 Variable selection techniques

Variable selection methods play a crucial role in statistical modelling and machine learning, as they facilitate the identification of relevant variables from the input set. These methods are crucial for several reasons and purposes [12]. Variable selection techniques encompass a wide range of approaches, including filter methods, wrapper methods, embedded methods and regularization methods [13]. Each method has its strengths and limitations, depending on the specific characteristics of the data set and the objective of the research.

Embedded methods are characterized by directly integrating the variable selection process into the model fitting procedure. These methods employ regularization techniques to simultaneously estimate the model parameters and select the most influential variables. By imposing penalties on the coefficients, embedded methods encourage sparsity and automatically identify the most important variables. Examples of such models are Lasso and Elastic Net [14, 15], described in the following section.

Filter and Wrapper algorithms use an external evaluation criterion to assess the quality of different subsets of variables. These algorithms typically involve a search procedure that evaluates the performance of various variable combinations using a selected model. Wrapper algorithms can be computationally intensive, as they repeatedly fit models on different subsets of variables. Examples of wrapper algorithms include forward selection, backward elimination and stepwise regression. These methods provide flexibility in terms of the evaluation criterion and allow for more customized variable selection based on specific objectives or domain knowledge [16].

While embedded methods gained popularity due to their simplicity and efficiency, wrapper algorithms provide more flexibility in the selection criteria and allow for fine-tuning the selection process based on specific requirements. Both criteria have their advantages and limitations and the choice between them depends on the nature of the data, the research objectives and computational considerations.

Let us consider a set of input variables denoted as X, with dimensions $n \times p$ where $n, p \in \mathbb{N}$. Here, n represents the number of observations and p represents the number of variables. We also consider the output variable, denoted as Y, with dimensions $n \times 1$.

3.2.1 Embedded methodologies

This section provides a brief overview of the two most popular embedded variables selection models, Lasso and Elastic Net.

3.2.1.1 Lasso

The Least Absolute Shrinkage and Selection Operator, also known as the Lasso model, is a linear regression model with regularization [14]. It incorporates a penalty term

that encourages sparsity by shrinking the coefficients of less important variables towards zero. It can be used as an effective approach for variable selection and regularization, allowing the identification of the most relevant variables from a set of input variables. This characteristic makes it particularly advantageous when dealing with high-dimensional data sets or in cases where the number of variables exceeds the number of observations.

The Lasso method aims to minimize the following objective function:

$$\min_{\beta} \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{X}_i \cdot \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{3.1}$$

where β denotes the coefficient vector and λ is the parameter that controls the degree of regularization. The first term in the objective function represents the residual sum of squares between the observed and predicted values. The second term is the penalty term that shrinks the absolute values of the coefficients towards zero encouraging sparsity. It aims to minimize the residual sum of squares, subject to the constraint that the sum of the absolute values of the coefficients is less than a specified tuning parameter λ. The tuning parameter λ plays a crucial role in controlling the degree of regularization applied. A large lambda value leads to more aggressive shrinkage, resulting in more coefficients being driven towards zero and, consequently, more variables being excluded from the model. Conversely, a small lambda value allows for less shrinkage and retains more variables in the model. The λ parameter is often tunned by cross-validation.

3.2.1.2 Elastic net

The Elastic Net model is a statistical technique that combines the advantages of both ridge regression and the Lasso method [15]. It is designed to address the limitations of each approach and provide a more robust framework for variable selection and regularization in linear regression models. The Elastic Net model introduces two penalty terms: the L1 penalty, which encourages sparsity by promoting some coefficients to exactly zero and the L2 penalty, which encourages shrinkage towards zero without enforcing strict sparsity. This combination allows for the selection of important predictors while handling multicollinearity among the variables.

The Elastic Net model aims to minimize the following objective function:

$$\min_{\beta} \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{X}_i \cdot \beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 \right) \tag{3.2}$$

where β denotes the coefficient vector, λ_1 controls the degree of L1 penalty (Lasso term), and λ_2 controls the degree of L2 penalty (Ridge term). A higher λ_1 value encourages more coefficients to be exactly zero, resulting in a restricted model. A

higher λ_2 value increases the shrinkage effect, allowing for better handling of multicollinearity among the predictors. By adjusting the values of λ_1 and λ_2, sparsity and shrinkage in the model can be controlled.

3.2.2 Wrapper based methodologies

In the following section will be presented briefly the Knockoffs variable selection [17, 18] and the Variable Selection Using Random Forest (VSURF) [19, 20].

3.2.2.1 VSURF

The VSURF (Variable Selection Using Random Forests) model is a wrapper built around the Random Forest (RF) algorithm [21]. It is specifically designed for variable selection to pursue two main objectives: interpretation or prediction. It aims to identify the most important variables from a set of input variables by iteratively fitting Random Forest models and evaluating each variable's importance score.

The VSURF algorithm is an iterative algorithm and is composed of essentially two steps. First, the algorithm starts by fitting the RF model using all available variables, running 50 runs of RF. Then, it evaluates the variable importance measures provided by RF, such as the mean decrease in accuracy (MDA) or mean decrease in impurity(MDI). Based on these measures, the least important variable is removed from the set. The elimination is based on an initial threshold set to the minimum standard deviation of the prediction value of the variable importance using a CART model estimation.

The second step depends on the objective of the variable selection. For interpretation purposes, the algorithm construct a nested collection of RF models involving the remaining variables from the previous step. It selects the variables involved in the model leading to the smallest OOB error. For prediction purposes, the algorithm build more parsimonious models. It starts from the ordered variables retained for interpretation and step-wise add a variable to the model only if the OOB error decreases significantly. The oob error is compared to the average oob error from the first step.

By iteratively eliminating the least important variables, VSURF gradually identifies the subset of variables that contribute most to the model's predictive accuracy. This allows for reducing model complexity and improving interpretability. The model presents other advantages. It can deal with cases when the number of variables is much higher than the observations, $p >> n$. It can pursue two objectives, so this makes it open to broad cases. This flexibility enables researchers to adapt the model to their specific needs and objectives. It uses RF as its base model, so VSURF inherits the RF properties as robustness to various data characteristics, such as high-dimensional data sets, multicollinearity and noisy data. On the downside, the VSURF model is computationally intensive as it iteratively fits multiple Random Forest models. It is not very suitable for cases that present high correlation among

the input data as it makes use of the MDA and MDI variable importance measure, which was demonstrated to be biased for highly correlated data [10]. It is important to consider these pros and cons while applying the VSURF model, as they can help researchers make informed decisions regarding variable selection approaches and their implications for the resulting model.

3.2.2.2 Knockoffs variable selection

The Knockoffs variable selection method is a statistical technique that aims to identify relevant predictors while controlling the false discovery rate (FDR) [22–24]. It provides a formal framework for variable selection by constructing a set of "knockoff" variables that mimic the structure of the original predictors.

Consider a subset of the input variables called S that affects the output variable Y. The subset of variables that do not influence the Y is called $\mathcal{H}_0$. Finally, let us suppose that the algorithm selects the variables included in the set $\hat{S}$. The False Discovery Rate can be defined as:

$$FDR = \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{max\{1, |\hat{S}|\}}\right]$$

The False Discovery Rate (FDR) is a crucial index within the context of the Knockoffs variable selection method. It serves as an initial estimate of the algorithm's efficacy in identifying relevant variables. The FDR represents the proportion of falsely selected variables among the ones claimed to be significant. When the FDR is low, it indicates that the algorithm has made fewer incorrect selections, implying a higher degree of accuracy in identifying relevant variables. Conversely, a substantially high FDR suggests that a large number of variables have been incorrectly identified as significant, indicating potential weaknesses or limitations in the algorithm's performance. Therefore, the FDR plays a crucial role in assessing the reliability and effectiveness of the Knockoffs method for variable selection tasks.

The algorithm proceeds as follows. The first step in the Knockoffs method is to generate a set of knockoff variables, denoted as $\tilde{X}$, which mimic the distributional properties and pairwise associations of the original variables X. The knockoff variables serve as a reference for evaluating the importance of the original variables while accounting for their correlation structure. Next, a statistical test is performed to compare the importance of each original variable to its corresponding knockoff variable. This test measures the difference in importance scores between the original and knockoff variables and provides a p-value associated with each variable. The p-values are then adjusted for multiple testing using a method such as the Benjamini-Hochberg procedure to control the FDR. Based on the adjusted p-values, variables with low p-values are deemed significant, indicating that they have higher importance than their knockoff counterparts. These variables are selected as the relevant variables.

The Knockoffs method offers several advantages. It provides a formal framework for variable selection and controls the FDR. It considers the correlation structure

among the variables and constructs knockoff variables that mimic the relationships among the original variable and it also accounts for potential confounding effects. Additionally, it allows for flexible choices in constructing the knockoff variables, such as using different variable transformations or incorporating prior knowledge.

However, it is important to note that the Knockoffs method has certain limitations. It assumes that the conditional distribution of each knockoff variable can be accurately estimated, which may be challenging in certain scenarios and bias variable importance measures. Additionally, the method may be computationally intensive, especially for high-dimensional data sets with a large number of variables. The target FDR rate must be defined at the beginning of the algorithm. In certain situations, an exceedingly low index can result in either no variable selection or the selection of only a minimal number of variables. When no variables are selected, two potential scenarios may arise: either no input variables are truly relevant, implying that only noise variables exist, or the index value is excessively low and requires adjustment. However, determining the appropriate extent by which the index should be increased remains an open question. Finding the optimal balance between avoiding false selections and accurately identifying relevant variables poses a challenge in this context. Considering these pros and cons is essential when applying the Knockoffs variable selection method. Understanding its strengths and limitations can help make informed decisions and interpret the results accurately in a specific research context.

3.3 Paper: Variable selection wrapper in presence of correlated input variables for random forest models

Marta Rotari and Murat Kulahci, *"Variable selection wrapper in presence of correlated input variables for random forest models"*, Quality and Reliability Engineering International (2023).

Received: 1 December 2022 | Revised: 2 June 2023 | Accepted: 7 June 2023

DOI: 10.1002/qre.3398

SPECIAL ISSUE ARTICLE

WILEY

Variable selection wrapper in presence of correlated input variables for random forest models

Marta Rotari[1] | **Murat Kulahci**[1,2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark

[2]Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

Correspondence
Marta Rotari, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark.
Email: mrot@dtu.dk

Abstract

In most data analytic applications in manufacturing, understanding the data-driven models plays a crucial role in complementing the engineering knowledge about the production process. Identifying relevant input variables, rather than only predicting the response through some "black-box" model, is of great interest in many applications. There is, therefore, a growing focus on describing the contributions of the input variables to the model in the form of "variable importance", which is readily available in certain machine learning methods such as random forest (RF). Once a ranking based on the importance measure of the variables is established, the question of how many variables are truly relevant in predicting the output variable rises. In this study, we focus on the Boruta algorithm, which is a wrapper around the RF model. It is a variable selection tool that assesses the variable importance measure for the RF model. It has been previously shown in the literature that the correlation among the input variables, which is often a common occurrence in high dimensional data, distorts and overestimates the importance of variables. The Boruta algorithm is also affected by this resulting in a larger set of input variables deemed important. To overcome this issue, in this study, we propose an extension of the Boruta algorithm for the correlated data by exploiting the conditional importance measure. This extension greatly improves the Boruta algorithm in the case of high correlation among variables and provides a more precise ranking of the variables that significantly contribute to the response. We believe this approach can be used in many industrial applications by providing more transparency and understanding of the process.

KEYWORDS

additive manufacturing, Boruta algorithm, conditional importance, random forest, variable selection algorithm

1 | INTRODUCTION

In many industrial applications, the production machines are equipped with several sensors in each stage of the production process, resulting in large amounts of data being collected. It is crucial to determine which variables from the process data actually influence the response variable, which is typically related quality of the final product. In this context, the interest is focused on the development and use of Machine Learning methods aimed at identifying the contribution of each process variable to the model, also known as variable importance, and in identifying the all-relevant variables, also known as variable selection, which play a crucial role in the mechanism of learning algorithms for predicting the response variable.

Variable selection is a challenging issue in high-dimensional regression and classification problems. Most of the input variables are usually irrelevant, and their relevance is unknown in advance. Typically, there are three goals in variable selection. The first is purely technical; dealing with large sets of variables slows down algorithms, consumes excessive resources and is inefficient.[1,2] A second aim is to find a small number of variables that maximize the model accuracy.[3] Numerous machine learning algorithms show a reduction in accuracy when the number of variables is significantly higher than optimal.[4] Therefore it is preferable to select the smallest set of variables that yields the best model. This problem, known as the minimal-optimal problem, has been explored extensively.[5]

The third aim is to improve understanding of the underlying process that generated the data.[6] The identification of all variables, which are in some circumstances relevant for classification or regression purposes, is the so-called all-relevant problem. The goal is to identify and prioritize all influential variables for further investigation with domain expertise. This is especially important when the goal is to understand mechanisms related to the subject of interest rather than simply building a black box predictive model. In the biomedical research, for example, when dealing with the results of gene expression measurements in the context of particular diseases, identifying an influential set of genes as genetic markers might be useful. In manufacturing, detecting all relevant variables of the production process could be very pertinent to understand and have a better overview of the process. It could complement the engineering knowledge about the production process and facilitate process improvement effects. Further details on the relevance of variable selection are described in Nilsson et al.[5]

Variable selection gets considerably more challenging in the presence of strongly correlated input variables, as it is common in real-world data. Consider two relevant variables which are strongly correlated. The second and third aims may seem similar but lead to different outcomes depending on which of these two objectives is of interest. The variables convey the same statistical information, so only one should be chosen if the goal is to maximize predictive accuracy with a small number of variables, second aim. On the other hand, these two variables may be collected in different ways and represent distinct physical quantities. Consequently, domain experts may interpret them differently, hence both should be preserved.

In machine learning for variable selection in the presence of high-dimensional data, Random Forest (RF)[7] model has been commonly used as a modeling method. The RF model consists of a collection of trees created using bagging or bootstrap aggregation. It is a nonparametric model for classification and regression issues. It has been applied to a wide range of problems, including high-dimensional problems with multi-class responses, categorical variables, imbalanced data and multiple adapted versions have been proposed.[8–12] The model is widely used due to its capacity to internally consider the interaction between input variables while providing variable importance measures. For variable importance in RF models, two types of measures have been proposed: the Mean Decrease Accuracy (MDA) and the Mean Decrease Impurity (MDI). Both measures are non-parametric and there have been several studies considering the theoretical formalization of these methods.[13–17] Nonetheless, further studies have been conducted on the influence of correlation on both variable importance measures,[18–21] which conclude that correlation among the input variables leads to a significant overestimation of the importance scores. According to simulation studies carried in literature,[18,22,23] highly correlated variables can erroneously show high MDA scores even when there is no dependence between the response variable and these variables. The MDA may fail to detect some relevant variables in the presence of correlation among the variables.[20,21,23–26]

While RF provides variable importance values, there is no built-in solution for variable selection based on a variable importance threshold. In the industrial context, the final decision on the subset of selected variables is manifest for the complete understanding of the processes. Moreover, a precise threshold is crucial to identify the most important variables without discarding any relevant information. Many techniques have been suggested on how to discard non-significant variables.[4,17,25,27] The most successful algorithms for this purpose are the wrapper methods,[24,28–33] which return a final

subset of all-relevant variables. To efficiently use a wrapper method, the model to be used should be both computationally efficient and simple, with no user-defined parameters if possible, which is the case for the RF models.

Boruta algorithm[34] is a wrapper method that aims to identify a clear threshold in the variable importance ranking provided by the RF model. It uses the MDA and MDI importance measures. Recently, a novel Python implementation of Boruta has been proposed, allowing the selection of any Tree-based learner.[35] Such importance measures quantify the relevance of an input variable towards a response variable by perturbing the values of the former. However, these do not consider the correlation among the inputs, therefore, the performance of the Boruta algorithm is compromised when correlated input variables are present. In particular, the problem of correlation leads to overestimated variable importance values. This overestimation can result in an incorrect ranking of the variables, producing a larger set of input variables deemed important.

In order to adjust the variable selection result in the presence of correlated input variables, we present an extension of the Boruta algorithm to effectively exploit the advantages of RF models with wrapper variable selection methods. This extension uses a conditional variable importance measure. The proposed algorithm significantly improves the variable importance ranking that results in a more precise final variable selection in comparison to the existing Boruta algorithm. In the following, we present the proposed extension of the Boruta algorithm and application on simulated dataset and on a real-world case in additive manufacturing to highlight the advantage of the current proposal.

2 | METHODS

2.1 | Model-based variable importance and variable selection

The first category of variable selection methods are modeling techniques that have the selection already built-in. The two most predominant in linear regression are Lasso and Elastic Net.[36,37] These methods make use of regularization to provide a measure of variable importance. The Lasso method includes a penalty term in its optimization criterion, which constrains the size of the estimated coefficients. As the penalty increases, the coefficients with the lowest values are set to zero. Similar to Lasso, Elastic Net can build reduced models by producing zero-value coefficients. Both methods produce reduced models with only the relevant variables.

Some Machine Learning methods, such as RF,[7] have embedded variable importance measures. Specifically, there are two commonly employed variable importance measures: mean decrease in impurity (MDI) also called the Gini importance and mean decrease in accuracy (MDA).[7] The former relies on the concept of impurity reduction followed in most traditional classification tree algorithms. First, the weighted impurity decrease over all nodes that split on a given variable are summed. The value is then averaged over all trees in the forest to obtain the MDI value. Furthermore, other analyses as in literature[26,38] have highlighted that MDI is consistent only under a strong and restrictive assumption of an additive regression function and independence of the input variables. Strobl et al.[26] claim that the MDI variable importance measure is biased when input variables vary in their number of categories or scale of measurement. On the other hand, MDA, also called permutation importance, is widely considered a more efficient variable importance measure for RFs.[13,26] This method can be more computationally intensive than MDI, but it may provide a more accurate estimate of variable importance when there are interactions among variables or when correlation among variables is present.[39] This paper focuses on regression settings and correlated data; therefore, the MDA measure will be considered.

Consider the input variables $X = (X_1, .., X_j, .., X_p)$ and the response variable Y. In the RF algorithm, each tree $t \in 1, \dots, n_{tree}$ is built with a bootstrap sample of the original data, $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ where $X_i = (X_{i1}, \dots, X_{ip})$. The left-over data from the bootstrap sample represents the out-of-bag (OOB) data, that we denote by $\mathbb{B}^{(t)}$. The OOB sample is used to evaluate the model prediction performance, also referred to as OOB error:

$$\hat{err}(\hat{f}^{(t)}, \mathbb{B}^{(t)}) = \frac{1}{|\mathbb{B}^{(t)}|} \sum_{i:(X_i,Y_i)\in\mathbb{B}^{(t)}} (Y_i - \hat{f}^{(t)}(X_i))^2 \tag{1}$$

where $\hat{f}^{(t)}(X_i)$ is the prediction for observation i and $|\cdot|$ is the cardinality function. To compute the MDA for variable X_j for a single tree, the values of variable X_j are permuted, yielding $\mathbb{B}^{(t,\pi_j)}$, following permutation π. The OOB error is computed again. The difference between the OOB error for the new data and the original OOB error is defined as the importance of the j-th variable for the tree t. The average of this difference over all trees in the forest constitutes the MDA

importance value for variable X_j and can be written as:

$$I(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left(\hat{err}(\hat{f}^{(t)}, \mathbb{B}^{(t,\pi_j)}) - \hat{err}(\hat{f}^{(t)}, \mathbb{B}^{(t)}) \right) \tag{2}$$

It is worth noticing that various MDA implementation exists. In standard RF implementations, an additional version of the permutation importance (often referred to as the z-score) is obtained by dividing the importance by its standard error. Bénard et al.[14] provide an exhaustive analysis of the various MDA implementations. Besides providing a more rigorous formalization, the authors propose a new, augmented version called Sobol-MDA that offers improved reliability and accuracy in measuring the variable's importance scores.

The random permutation of the X_j values breaks the initial relationship between the independent variable Y and the dependent variable X_j. It should simulate the lack of the variable in the model. If the original variable X_j is associated with the response variable, the OOB error increase when the permuted variable X_j and the other non-permuted input variables are used to predict the response for the OOB observations. The difference between the accuracy of prediction before and after permuting the values of X_j can be viewed as a measure of the importance of X_j in predicting the response variable Y. When there is almost no difference in the accuracy of the forecast before and after permutingX_j, X_j is said to be unimportant.[18,25,26]

2.2 | Variable selection based on the Boruta algorithm

The Boruta algorithm[34,35,40] is a wrapper algorithm developed for the RF model. In a wrapper method, an algorithm is used as a black box that returns a variable ranking. The RF regression and classification algorithm is relatively quick, can usually be run without tuning any parameters, it is sensitive to the interaction between variables without explicit settings and gives a numerical estimate of the variables' importance. The Boruta algorithm is a sequential selection algorithm, and each step consists in running RF and obtaining the variables' importance values. It offers a precise threshold in the RF variables ranking in order to decide a final set of relevant variables to predict the output variable.

The algorithm is an extension of the original proposal in literature,[17] that determines the relevance of a variable by comparing the importance of the original variables to that of the random noise variables. Based on the values obtained from the RF model, Boruta evaluates which variables are relevant. The reference for deciding which variables are truly relevant is given by the set of the noise variables. The main goal is to find all variables for which their association with the response variable is higher than that of the noise variables. Numerous RF realizations in the Boruta algorithm produce a more stable output than a single RF run. The original algorithm description can be found in literature.[34,40]

2.3 | The influence of the correlation on the permutation importance measure

Strobl et al.[18] formalize the MDA interpretation under the assumption of no correlation among input variables and the response variable Y. They argue that if there is independence between the variable X_j and the response Y and marginal independence between X_j and the other variables $X \setminus X_j$, then the permutation of X_j would not affect the prediction accuracy. An MDA importance value close to zero validates the hypothesis of marginal independence. Consequently, a large value can be indicative of dependence between X_j and Y or between X_j and other variables, or both.

The effect of the correlation among input variables on the MDA measure and its bias has been studied in the literature.[18–21] However, there is no agreement on how to interpret the importance measures when the input variables are correlated and even less agreement on how this correlation affects the importance measures.[41,42] Nicodemus et al.[22] show through simulated studies that highly correlated variables acquire high MDA values even when there is no dependence on the response variable. Strobl et al.[18] highlighted two issues in the high MDA values of correlated variables. The first reason is identified in the tree-building process that prefers the selection of correlated variables. The second is identified in the computation of the MDA value and the advantage of the correlated data induced by the unconditional permutation scheme. Toloşi and Lengauer[23] identify this effect as "correlation bias", which does not correspond to a statistical bias. They observe a critical effect of the correlation on the permutation importance measure that depends on the size of the

correlated group. Similarly, other empirical studies show that MDA fails to identify some of the relevant variables when highly correlated variables are present.[20,21,23–26]

Following that, Gregorutti et al.[25] provides a more formal description and proof of this effect, limited to a regression setting. In the case of an additive regression model, it is possible to express the permutation importance measure as a function of the correlation between input variables. Consider (X, Y) random vectors which satisfy the following additive regression model

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{3}$$

where ϵ is such that $\mathbb{E}[\epsilon|X] = 0$ and f_js are measurable functions. In this model setting, for any $j \in 1, \dots, p$ the permutation importance measure satisfies

$$I(X_j) = 2\mathbb{V}[f_j(X_j)]. \tag{4}$$

That is, the variable importance score of the X_j variable is equivalent to two times the variance of $f_j(X_j)$. Assume moreover that for some $j \in 1, \dots, p$ the variable $f_j(X_j)$ is centered. Then

$$I(X_j) = 2\mathbb{C}[Y, f_j(X_j)] - 2\sum_{l \neq j} \mathbb{C}[f_j(X_j), f_l(X_l)]. \tag{5}$$

where $\mathbb{C}$ denotes the covariance. Note that Equation (5) reveals the strong dependence on the additive structure of the regression function f.

Even if restricted to special model settings, the above equations describe how the correlation among the variables impacts the MDA. Therefore the correlation among the variables is to be considered when interpreting the variable importance values in RF models. Furthermore, high correlation among input variables leads to a considerable overestimation of the variable importance values of the correlated variables. Consequently, all correlated but non-influencing variables will be highly ranked in importance. The Boruta algorithm is also highly affected by this effect as RF and in particular its measures are the fundamental of Borutas' algorithm. The algorithm deem a large number of variables to be relevant. The majority of the selected variables are variables with a high empirical correlation. As a consequence, some irrelevant variables might be deemed relevant and some variables with low relevance might be missed. We next present the extension of the Boruta algorithm involving the correlation among the input variables.

2.4 | Extension of Boruta algorithm to the case of high correlated input variables

2.4.1 | The use of conditional variable importance

Strobl et al.[18] propose a new importance measure for RF models that is based on a conditional permutation scheme that better reflects the impact of input variables on the response variable when a high correlation among the input variables is present. The aim is to evaluate the deviation from the null hypothesis, that X_j and Y are independent based on the correlation structure between X_j and the other variables.

The strategy of this new measure is to build a conditional permutation scheme in the dataset based on the correlation among variables, with the aim to preserve the data correlation structure. To calculate the variable importance of a variable X_j, the values of X_j are permuted based on the so-called conditional permutation scheme. The nature of conditional permutation comes from the fact that X_j is permuted only within groups of observations with $Z^{(j)} = z$, where $Z^{(j)}$ represent the set of variables that are correlated with X_j. The exact variables that should be included in the subset $Z^{(j)}$ are those whose correlation with X_j exceeds a certain threshold. Another option is to allow the user to specify which variables to condition on, for example, if a hypothesis of interest contains certain independencies.

We present here the derivation of the conditional mean decrease in accuracy (CMDA). Consider the input variables $X = (X_1, \dots, X_j, \dots, X_p)$ and the output variable Y. First, the calculation of CMDA is done for every tree of the RF model.

In every tree in the model $t \in 1, \dots, n_{tree}$ the predictions for the OOB data $B^{(t)}$ are calculated. Next we proceed with the following construction of the conditional permutation scheme:

1. Select a subset of variables Z to base the conditional permutation on. The subset is composed of variables that show high correlation with variable X_j greater than a predefined threshold. $Z^{(j)} = \{X_i \mid cor(X_i, X_j) > threshold, X_i \in X \setminus X_j\}$
2. Collect the split points from the tree t of the RF model for $Z^{(j)}$.
3. Use the split points to create a grid that bisects the input variable space.
4. Within the obtained grid permute the values of X_j and compute the OOB-prediction error again.

The value of CMDA for variable X_j for one tree corresponds to the difference in prediction accuracy before and after this conditional permutation. We then repeat the procedure for every tree in the forest ($t \in 1, \dots, n_{tree}$). The final CMDA value of variable X_j is given by the average over all CMDA values calculated from every tree as:

$$I(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left(\hat{err}(\hat{f}^{(t)}, \mathbb{B}^{(t,\pi_j|Z)}) - \hat{err}(\hat{f}^{(t)}, \mathbb{B}^{(t)}) \right) \tag{6}$$

where $\hat{err}(\cdot, \cdot)$ is defined in (1).

In the presence of strongly correlated variables, this new technique is able to determine the variable importance values more accurately. In fact, CMDA limits the variable importance values of correlated data; thus, variables with a stronger impact on the response variable have a greater chance of being identified. Even if this approach fails to totally eliminate the influence of correlated variables, by employing this new variable importance measure technique, we can obtain a more accurate representation of the variables' relevance in relation to the response variable. Compared to MDA, this method has shown substantial improvement in circumstances of highly correlated data.[18]

2.4.2 | Extended Boruta algorithm

Using the CMDA importance measure, we present here the complete definition of the extended Boruta algorithm. Consider the input space $X_1, \dots, X_p$ and the response variable Y. In this algorithm the variables are classified as relevant, irrelevant and tentative. A variable that is classified as *relevant* is considered a selected variable. A variable that is classified as *tentative* is a variable neither selected nor rejected. Finally, a variable classified as *irrelevant* is a rejected variable. Just as in the case of the Boruta algorithm the noise variables are used as a benchmark for the importance of the input variables. The algorithm starts by marking all input variables as "*Tentative*" variables. The algorithm runs for a number of iterations until all the variables are classified or the maximum number of iterations is reached. At each iteration, the algorithm proceeds as follows:

1. The dataset is expanded by including noise variables. The noise variables are obtained by shuffling the values of at least five initial variables selected from variables marked as tentative.
2. With the expanded dataset an RF model is built.
3. The obtained RF model is used to calculate the CMDA value for each variable of the extended dataset, as explained in section 2.4.1.
4. The testing threshold is set. This threshold corresponds to the maximum of the CMDA values of the noise variables.
5. The CMDA values of each X_j variable is compared with the testing threshold. If the CMDA_j value exceeds the testing threshold, the variable X_j is given a score point of 1 or 0 otherwise.
6. The obtained scores for all input variables are added to the scores from the previous iterations. This produces a vector called hints $H = (h_1, \dots, h_j, \dots, h_p)$, $H \in \mathbb{N}^p$.
7. A classification of the input variables is made based on the hints vector H. The classification follows the Binomial decision scheme, discussed below.
8. If not all the variables are classified as relevant or irrelevant, a new iteration starts. Otherwise the algorithm ends.

The Binomial decision scheme is based on the definition of the hints vector. Each iteration of the algorithm is assumed to be an independent experiment that gives a binary outcome. The H random vector represents the number of success

in a binomial distribution, with n the number of iterations and $p = 0.5$. We assess the probability that an input variable scores better than the maximum noise variable importance value. The binomial decision scheme is taken from the original Boruta algorithm. Consider a binomially distributed random variable $B \sim Bi(n, 0.5)$ where n is the number of iterations and p is equal to 0.5. Using a significance level α, the decision proceeds as follows:

1. To classify variable X_j as *relevant* we evaluate if the obtained scores h_j:

$$P(B > h_j - 1) < \alpha \tag{7}$$

   where $P(B > h_j - 1) = 1 - P(B \leq h_j - 1)$. The obtained p-value is adjusted with the Bonferroni correction and then compared to the significance level α.
2. To classify variable X_j as *irrelevant*, we evaluate if:

$$P(B \leq h_j) < \alpha \tag{8}$$

   Once again, the p-value is adjusted by Bonferroni correction before comparing with α.
3. If a variable does not satisfy either of the above cases, the variable remains marked as *tentative* variable.

We recall that the Bonferroni correction is an adjustment for multiple comparison tests used in statistical analysis. When performing a hypothesis test with multiple comparisons, wrongly claiming statistical significance in at least one of these comparisons is higher than the $\alpha-$level set for each comparison. Considering a family of m hypotheses for significance testing and their corresponding p-values p_i with $i = 1, \ldots, m$, the Bonferroni correction rejects the null hypothesis for each $p_i \leq \alpha/m$.

When the maximum number of iterations is reached or all of the original variables are classified as relevant or irrelevant, the algorithm terminates. It may happen that a variable has not been classified at the end of the algorithm, that is, it remains as a *tentative* variable. The tentative variable has an importance score that is very close to the best noise variable importance value, and the Boruta algorithm is unable to make the desired decision in the default number of iterations. In this case, there are two options: increase the number of iterations or compare the median importance variable value with the maximum importance value for the noise variables. This is based on the historical variable importance stored in memory by Boruta for each iteration. For further details of these rules see.[40]

The newly introduced extension requires a longer computation time than the original version. The CDMA, as described in Strobl et al.,[18,43] offers a significant advantage when dealing with highly correlated data, but requires more computational time than other variable importance measures. Despite the computational cost, the benefits of using this measure are substantial, as discussed in the following section. The decision to employ this method should be based on the correlation among the variables and the research objectives.

The proposed method's R code is available online,[44] providing researchers with an accessible tool to implement the method. The R function allows for control over fundamental parameters: the significance level α, the maximum number of iterations, `mtry` the number of variables randomly selected at each split, and `ntree` the number of trees in RF. A brief discussion on how to choose these parameters can be found in the next section.

3 | RESULTS AND DISCUSSION

This section is devoted to evaluating the proposed model's effectiveness, which we demonstrate through two simulated datasets and an industrial case study. We also include a comparison with existing models such as the original Boruta, Lasso, Elastic net, Knockoffs variable selection and Variable Selection Using Random Forest (VSURF). The first two models, Lasso and Elastic Net, are two commonly used regression models with regularization. Knockoffs variable selection[45,46] is a variables selection model that uses a set of "knockoff" variables and it is designed to control for false discovery rate. VSURF[24,47] is also a variable selection method employing a RF-based approach.

The first simulation aims at assessing the model's performance in detecting the most significant variables among all variables in the presence of multiple groups of correlated variables. Additionally, this simulation includes a brief discussion of the model's parameters. The second simulation evaluates the model's performance when the interactions among variables are present. The utilization of simulated data in our evaluation process serves two purposes. Firstly, it allows

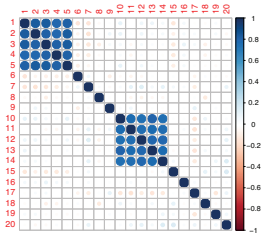


FIGURE 1 Correlation structure of the input variables.

for greater control over the correlation among the variables, enabling a more systematic investigation of the model's performance under varying conditions. Secondly, using simulated data provides a comparison of various model selection approaches. In section 3.2, we present a real-world case in additive manufacturing. The obtained results were evaluated by process engineers and compared with the existing knowledge about the process. All the analyses were performed in R language, version 4.2.3, using functions available at[44] and publicly available R packages.

3.1 | Simulated dataset

In the first simulation study, two groups of correlated variables are introduced in the dataset, and the simulations were run at increasing correlation levels from 0 to 0.9. The 20 input variables given in X are drawn from a joint normal distribution with mean 0 and variance 1. Two groups of correlated variables are introduced: $[X_1, X_2, X_3, X_4, X_5]$ and $[X_{10}, X_{11}, X_{12}, X_{13}, X_{14}]$, as shown in Figure 1. The response variable is generated as a linear combination of X' where $X' = [X_2, X_{11}, X_{19}, X_{20}]$ and the coefficients β_i for $i \in \{2, 11, 19, 20\}$ are such that $\beta_i / sd(\beta_i) = k$, where $sd(\cdot)$ is the standard deviation and $k = [4, 3, 3, 5]$. The model for Y is given as

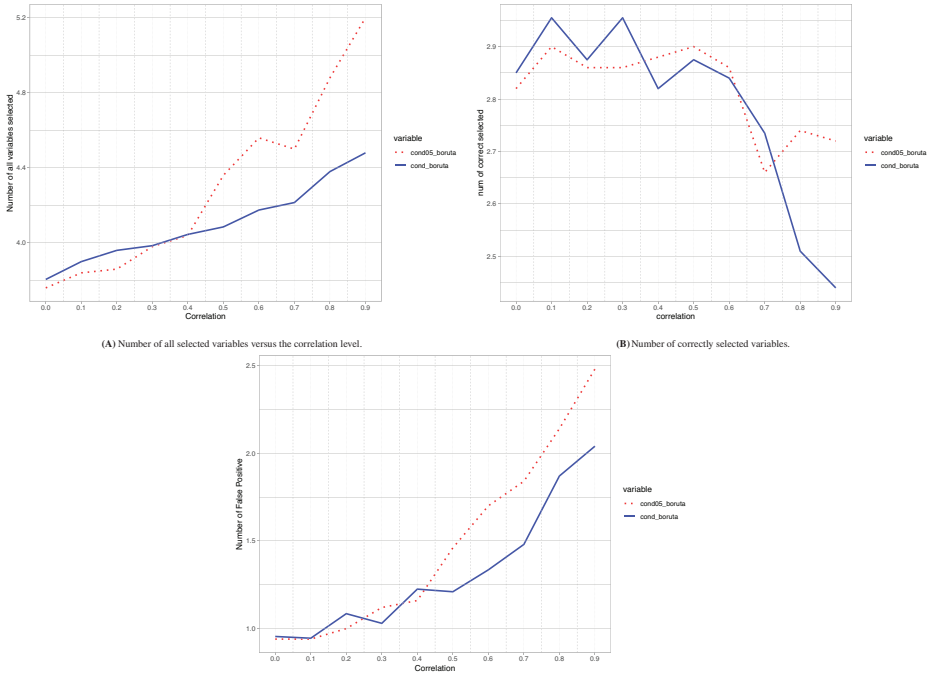
$$Y = \beta_2 X_2 + \beta_{11} X_{11} + \beta_{19} X_{19} + \beta_{20} X_{20} + \epsilon \tag{9}$$

where ϵ represents the added noise with $\epsilon \sim N(0, 0.1)$.

We performed an analysis to assess the model sensitivity to the choice of significance level α, the number of variables randomly selected at each split (`mtry`) and the number of trees in the RF (`ntree`). In each case we performed 50 simulation runs. The first parameter being investigated is the significance level α, for which the default value is 0.01. Figure 2 depicts a comparison between the default value of 0.01 against 0.05, a level that is also commonly used in regression.

Figure 2 demonstrates that the total number of selected variables is expectantly higher for $\alpha = 0.05$ compared to $\alpha = 0.01$. This implies that on average a greater number of significant variables are being correctly selected. However, an increase in the number of selected variables also results in the selection of noise variables that are not relevant to the response variable, Y. Careful consideration of the positive and negative effects associated with different α-levels is necessary to make an informed decision. Determining the appropriate significance level depends on the specific case and objectives of the analysis.

The sensitivity of the model to the changes in `mtry` is depicted in Figure 3. As discussed in literature,[18,43] the CMDA method is moderately sensitive to the choice of `mtry`, albeit less so than the MDA method. This phenomenon is because correlated variables are favoured in the split selection process. Consequently, for low values of `mtry`, correlated variables are more likely to be chosen, resulting in a higher variable importance score than the uncorrelated and noise variables. As can be seen in Figure 3A, a higher number of variables are selected for low `mtry` values. This increase is partly due to the selection of noise variables that are highly correlated with the relevant variables. As shown in Figure 3B, in the case of correlation level 0.8 and low `mtry` values, the entire correlation group associated with $X2$ is selected, in addition to $X2$ itself. This effect is less pronounced for the second correlation group, as $X11$ has a lower regression coefficient. The sensitivity of the CMDA method to the choice of `mtry` suggests that careful consideration of this parameter is necessary to ensure an appropriate model.

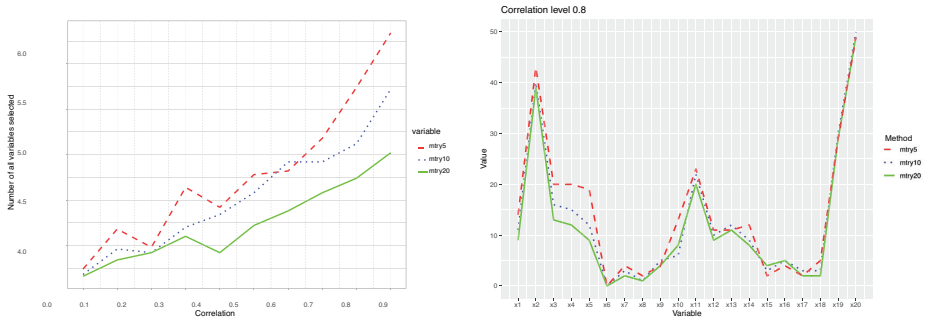


(A) Number of all selected variables versus the correlation level.

(B) Number of correctly selected variables.

(C) Number of wrongly selected variables.

FIGURE 2 In the solid blue line, the default value $\alpha = 0.01$ and the dotted red line $\alpha = 0.05$. (A) Number of all selected variables versus correlation. (B) Mean number of correctly selected variables over 50 runs versus correlation. (C) Number of wrongly selected variables as the correlation increases.



(A) Number of all selected variables versus the correlation level.

(B) Total number of selected variables for 50 runs.

FIGURE 3 In the solid green line `mtry = 20`, the dotted blue line `mtry = 10` and dashed red line `mtry = 5`. (A) Number of all selected variables versus correlation. (B) Number of times each variable has been selected in 50 runs.

We also assessed the impact of `ntree` parameter by evaluating the model's performance for `ntree = 500, 1000` and 2000 trees. We observed no substantial changes in the model's performance as the number of trees increased. Given that the proposed method is computationally more demanding compared to the original method, we recommend using the default value of `ntree = 500`. This choice strikes a reasonable balance between model accuracy and computational efficiency.

In the second simulation study, the data set is similar, except for the addition of an interaction term in the model. That is, the response variable is generated as a linear combination of X' where $X' = [X_2, X_7X_{11}, X_{11}, X_{19}, X_{20}]$ and the coefficients β_i is such that $\beta_i/sd(\beta_i) = k$, where $sd(\cdot)$ is the standard deviation and $k = [4, 5, 3, 3, 5]$. The model for Y is

$$Y = \beta_2 X_2 + \beta_{7,11} X_7 X_{11} + \beta_{11} X_{11} + \beta_{19} X_{19} + \beta_{20} X_{20} + \epsilon \tag{10}$$

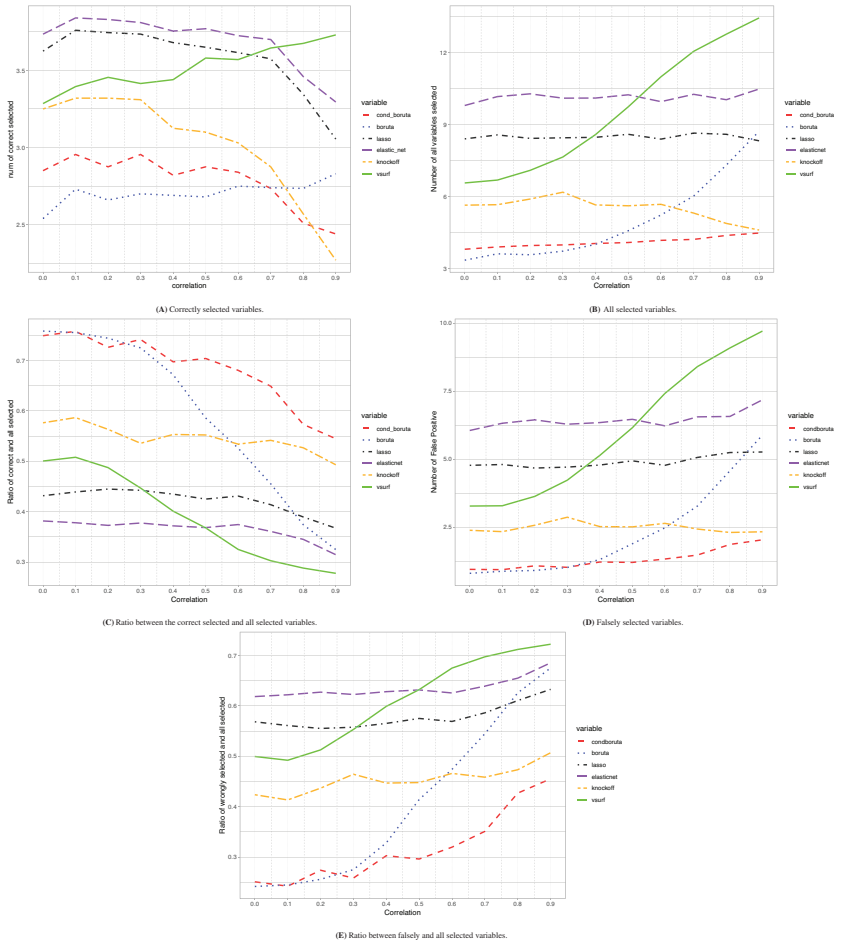
with ϵ represents the added noise with $N(0, 0.1)$. At each correlation level, the proposed model as well the original Boruta, Lasso, Elastic Net, Knockoffs and VSURF models were run 200 times each. The Lasso and Elastic Net model parameters were estimated through five-fold cross-validation. We show the model comparison results from the first simulation dateset for illustration purposes in Figure 4. The other simulation runs show similar results and hence omitted.

In both simulation studies, the Elastic Net and Lasso models gave similar performances. Both methods, on average, select one more correct variable than the other models, as can be seen in Figure 4A. However, Figure 4B shows that both models on average select a larger number of variables even for low correlation among variables. The number of selected variables is more than double the number of truly significant ones. Yet so, the number of the selected variables is nearly constant for increasing levels of correlation. This is further supported by Figure 4C, which shows that the ratio between variables correctly selected and the total number of selected variables related to the correlation level generally remains the same. The number of wrongly selected variables shown in Figure 4D, is stable for increasing correlation levels and this trend is also reflected in the ratio between wrongly selected and all selected variables as shown in Figure 4E.

The Knockoffs variable selection method results are also presented in Figure 4. One of the key parameters of the model is the target false discovery rate (FDR), with a default value of 0.1. However, with this setting, only 0.5% of the models selected any variables, meaning that 99.5% of the models were empty and had selected 0 variables. Therefore, we had to increase the value of the FDR to 0.5 to ensure that at least one variable was selected in at least 90% of the models. We observed that the number of models with no terms increases proportionally with the correlation levels. In other words, only a few models contained no terms for low correlation levels, however for high correlation levels, an increasing number of models ended up with no terms. Therefore we conclude that, as the correlation between variables increases, the method may become less effective in identifying relevant variables and may result in an increasing number of cases with no terms being selected. Figure 4B shows that the Knockoffs model selects a constant number of variables. This is also reflected in the ratio between correctly selected and all selected variables shown in Figure 4C, as well as the ratio between wrongly selected and all selected shown in Figure 4E. These ratios remain relatively constant but at a lower level than for the proposed model.

In Figure 4 we also present the results for the VSURF model. We notice that as the correlation increases, the number of selected variables rises drastically, negatively impacting the ratio in Figure 4C. As correlation increases, the number of incorrectly selected variables also increases, indicating a growing number of noise variables being selected. In Figure 4E, we can notice that more than half of the variables selected are noise variables.

We also observe a difference between the original Boruta algorithm and the proposed extension. On average, the extended Boruta algorithm correctly selects a greater number of significant variables. We can also see that as the correlation increases, the total number of variables selected by the original Boruta grows quickly. Instead, even in the case of high correlation, the proposed extension maintains a constant number of selected variables as shown in Figure 4B. This is reflected, also, in the ratio between correctly selected variables and all selected variables. The conditional Boruta selects a large number of significant variables while keeping the total number of variables selected at a low level. This is also supported by the ratio of correctly chosen variables to the total number of variables selected, shown in Figure 4C. Moreover, the proposed model exhibits a very low number of wrongly selected variables in Figure 4D, in fact the lowest among all the models studied in this work. The ratio between wrongly selected and all selected variables is also the lowest among all models. These findings suggest that the selected variables accurately reflect the variables used to construct Y, even in challenging scenarios with high correlation among variables for which other models tend to over select variables.



(A) Correctly selected variables.

(B) All selected variables.

(C) Ratio between the correct selected and all selected variables.

(D) Falsely selected variables.

(E) Ratio between falsely and all selected variables.

FIGURE 4 (We present the results from only one of the simulated data sets for illustration.) (A) Number of correctly selected variables by six different methods versus the correlation among the variables. (B) Number of all selected variables versus the correlation among the variables. (C) The ratio between the correctly selected and all selected variables versus the correlation among the variables. (D) Number of falsely selected variables versus the correlation among the variables. (E) The ratio between falsely selected and all selected variables versus the correlation among the variables.

In Figure 5, we present the frequency of each variable being selected by different models. The input variables are displayed on the x-axis of each figure, and the y-axis represents the frequency with which each variable was selected by the algorithms across 200 iterations. We can see that all models, on average, pick the relevant variables correctly. However, even at low levels of correlation among variables, Lasso, Elastic Net and VSURF also select noise variables as significant. As the correlation increases, the original Boruta algorithm selects the entire group of correlated variables. In general, we also notice a difference in the frequency of selection of the two correlated variables $X2$ and $X11$. The variable $X11$

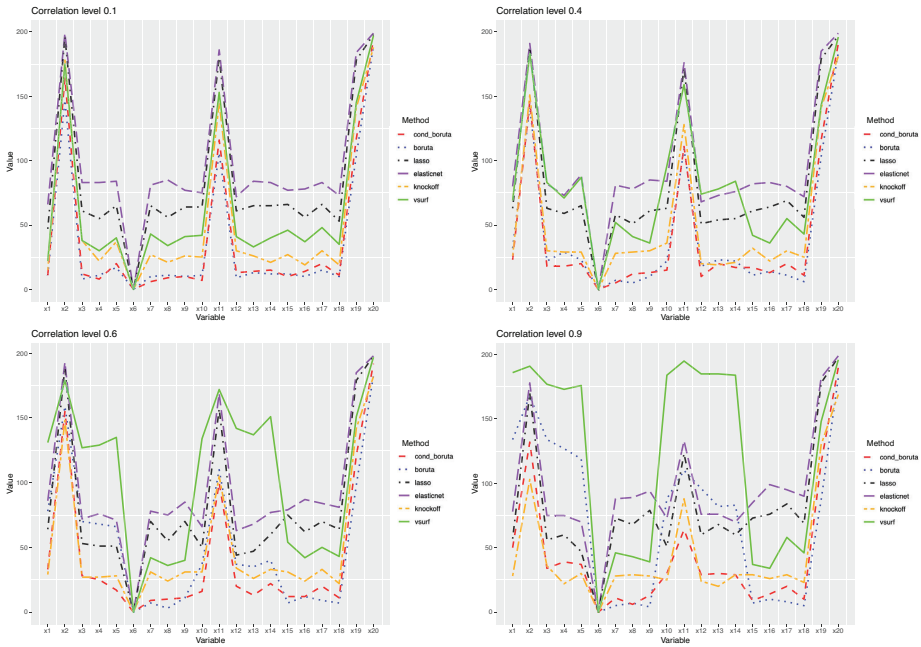


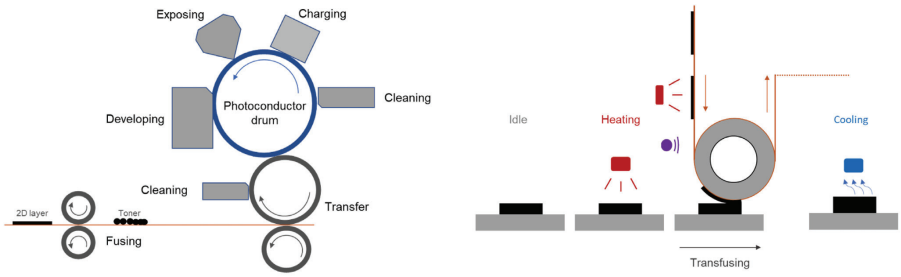
FIGURE 5 (We present the results from only one of the simulated data sets for illustration.) The frequency of the variables being selected in 200 runs for correlation levels: $cor = 0.1, 0.5, 0.7, 0.9$.

is selected less frequently than the variable $X2$, this is due to a lower coefficient used in (9). The same happens for the uncorrelated variables $X19$ and $X20$, for which the former is not selected in all models having a lower coefficient.

The Boruta algorithm, at each iteration, stores the variable importance value for the original variables and averages them at the end of the algorithm in a final table. The extension we propose in this paper gives us a more accurate selection of variables and a more representative final table. Further analysis could be carried out by exploiting the ranking of the variable's importance values in the final table, particularly the variables close to the Boruta threshold. This possibility can be pursued if Boruta's outcome is not entirely satisfactory. We can increase the number of variables or, on the contrary, be more restrictive in our decision by utilizing subject matter knowledge.

3.2 | Application to additive manufacturing case

In this section, we demonstrate the use of the proposed method for variable selection in an actual production data. The data is acquired from a brand new additive manufacturing equipment for high volume 3-D printing. The Selective Thermoplastic Electrophotographic Process (STEP)[48] is a breakthrough approach in additive manufacturing, that offers a very flexible option for complex geometries and various aspects of colouring. This new technology works by fusing and pressing super-thin, nearly two-dimensional layers produced by electrophotography into a single 3D bulk structure. With its two fundamental modules, electrophotographic and transfusion, Figure 6, this technology is able to produce a completely dense, multi-material, and multi-coloured components.[49] Multiple sensors are positioned throughout the production chain on the new manufacturing line. Examples of measured quantities include the amount of material used for each layer, the temperature before and after the melting process.



(A) Electrophotographic module.

(B) Transfusion module.

FIGURE 6 Additive manufacturing process: STEP. Adaptation from[48] of the Electrophotographic and Transfusion modules. STEP, Selective Thermoplastic Electrophotographic Process.

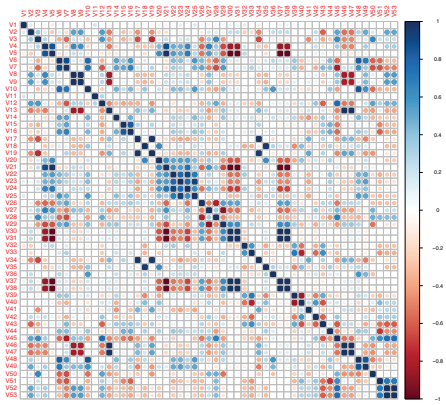


FIGURE 7 Additive manufacturing real-world application. Correlation among the input variables.

The main challenge is that the printing process is not very well understood and the goal is to identify key process variables that are related to product quality through data-driven approaches. The case study involves 18 different batches of production. As the input data, we considered 53 continuous variables collected through the sensors located throughout the printing machine. As the response variable, we consider Young's modulus as the physical quality aspect of the final products. Young's modulus is a fundamental concept in materials science that measures the stiffness of the material, and it is defined as the ratio of stress to strain in a material subjected to tensile or compressing forces. The input variables are labelled from $V1$ to $V53$ due to confidentiality.

Figure 7 displays the correlation among the production process variables, revealing differing degrees of correlation between the variables. Nonetheless we expect that particularly the high correlation, will likely cause certain variables' importance scores to be overestimated. This, in turn, could lead to the inaccurate ranking of variables and, therefore, to the selection of irrelevant variables for further studies.

All the models previously discussed in this paper were applied to the manufacturing data, showing similar prediction performances. Figure 8 shows the resulting variable importance rankings. The parameter estimates in Lasso and Elastic Net models are determined using five-fold cross-validation. We can see that these two models result in very similar

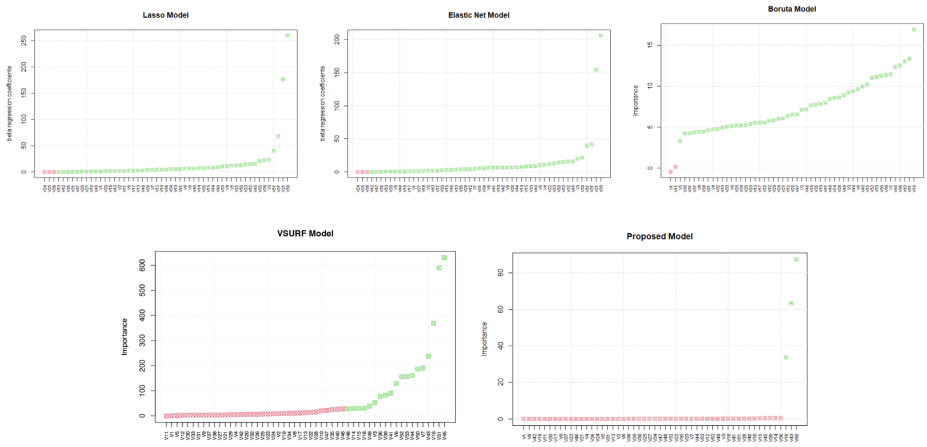


FIGURE 8 Lasso model, Elastic Net, Boruta, VSURF and the proposed model applied to the additive manufacturing real-world application. VSURF, Variable Selection Using Random Forest.

rankings of the variables. Both models select most of the variables, 50 out of 53 variables, to be relevant. The original Boruta algorithm selects most of the process variables, 51 out of 53, as relevant. The VSURF model selected 19 out of 53 variables, indicating a more conservative approach when compared to the first three models. Nonetheless, the selection of a relatively high number of variables is still noteworthy. The Knockoffs variable selection model provides only the selected variables. The outcome of this selection method shows $V3$, $V7$, $V10$, $V39$, $V40$, $V41$, $V45$, $V49$, $V50$, and $V51$ as the selected variables. This model presented a more restrictive selection compared to the aforementioned methods, and the selected variables align with the ones highly ranked and selected from the previous methods. The proposed model, on the other hand, is the only one that selects a significantly smaller number of variables, that is, three variables ($V16$, $V43$ and $V50$) out of 53 are selected as relevant. These variables correspond to distinct stages of the process. $V16$ is associated with the way the layers overlap. In fact, if the layers do not overlap properly, the final product's quality and, in particular, the physical characteristics may be compromised. $V43$ and $V50$ are connected to the layer positioning belt, which is subject to high degradation. The first, $V43$ is connected to the electrophotography module Figure 6A and it is related to the transfer of the image to the belt. Variable $V50$ tracks the number of hours the belt has been in operation. The engineers have observed that it becomes defiled after a specific number of production hours. Thus, additional investigation will determine the optimal number of hours to replace this component. With the engineers' approval, these variables are selected to be investigated further.

4 | CONCLUSION

In this study, we present an extension of the original Boruta algorithm for the case of high correlation among input variables. This extension makes use of the conditional variable importance measure, which is a more sensitive measure in the case of highly correlated variables. To evaluate the performance of the proposed extension, two simulation studies and a real-world case are presented. The results of the proposed extension are compared against other variable selection approaches, including the original Boruta algorithm, Lasso and Elastic Net regression, VSURF, and the Knockoffs variable selection method. Our findings indicate that the proposed extension outperforms these other methods in terms of identifying the most relevant variables while minimizing the number of wrongly selected variables, particularly when the correlation among variables is high and in the case of variable interaction. Moreover, the extended method also exhibits superior performance in terms of the ratio of correctly selected variables to the total number of selected variables. In the industrial case study, the proposed model selects fewer variables than other models that select most of the input variables.

We believe that this approach can be used in many applications, as it provides greater transparency and understanding of the process.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Marta Rotari https://orcid.org/0000-0001-5263-1937
Murat Kulahci https://orcid.org/0000-0003-4222-9631

REFERENCES

1. Kiziloz HE, Deniz A. An evolutionary parallel multiobjective feature selection framework. *Comput Ind Eng.* 2021;159:107481.
2. Shinde A, Church G, Janakiram M, Runger G. Feature extraction and classification models for high-dimensional profile data. *Qual Reliab Eng Int.* 2011;27(7):885-893.
3. Atamuradov V, Medjaher K, Camci F, Zerhouni N, Dersin P, Lamoureux B. Feature selection and fault-severity classification–based machine health assessment methodology for point machine sliding-chair degradation. *Qual Reliab Eng Int.* 2019;35(4):1081-1099.
4. Kohavi R, John GH,. Wrappers for feature subset selection. *Artif Intell.* 1997;97(1-2):273-324.
5. Nilsson R, Pena JM, Björkegren J, Tegnér J. Consistent feature selection for pattern recognition in polynomial time. *J Mach Learn Res.* 2007;8:589-612.
6. Detzner A, Eigner M. Feature selection methods for root-cause analysis among top-level product attributes. *Qual Reliab Eng Int.* 2021;37(1):335-351.
7. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5-32.
8. Segal M, Xiao Y. Multivariate random forests. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2011;1(1):80-87.
9. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS,. Random survival forests. *Ann Appl Stat.* 2008;2(3):841-860.
10. Conn D, Ngun T, Li G, Ramirez CM,. Fuzzy forests: Extending random forest feature selection for correlated, high-dimensional data. *J Stat Softw.* 2019;91:1-25.
11. Zhu M, Xia J, Jin X, et al,. Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access.* 2018;6:4641-4652.
12. Tang F, Ishwaran H. Random forest missing data algorithms. *Stat Anal Data Min.* 2017;10(6):363-377.
13. Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat.* 2007;1:519-537.
14. Bénard C, Veiga dS, Scornet E. MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika.* 2022;109:881-900.
15. Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat Med.* 2019;38(4):558-582.
16. Zhu R, Zeng D, Kosorok MR,. Reinforcement learning trees. *J Am Statist Assoc.* 2015;110(512):1770-1784.
17. Stoppiglia H, Dreyfus G, Dubois R, Oussar Y. Ranking a random feature for variable and feature selection. *J Mach Learn Res.* 2003;3:1399-1414.
18. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics.* 2008;9(1):1-11.
19. Archer KJ, Kimes RV,. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal.* 2008;52(4):2249-2260.
20. Nicodemus KK, Malley JD,. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics.* 2009;25(15):1884-1890.
21. Auret L, Aldrich C. Empirical comparison of tree ensemble variable importance measures. *Chemom Intell Lab Syst.* 2011;105(2):157-170.
22. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics.* 2010;11(1):1-13.
23. Toloşi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics.* 2011;27(14):1986-1994.
24. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett.* 2010;31(14):2225-2236.
25. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. *Stat Comput.* 2017;27(3):659-678.
26. Strobl C, Boulesteix AL, Augustin T. Unbiased split selection for classification trees based on the Gini index. *Comput Stat Data Anal.* 2007;52(1):483-501.
27. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003;3:1157-1182.
28. Svetnik V, Liaw A, Tong C, Wang T. *Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules*. Springer; 2004:334-343.
29. Louw N, Steel S. Variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination. *Comput Stat Data Anal.* 2006;51(3):2043-2055.

30. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn.* 2002;46(1):389-422.
31. Rakotomamonjy A. Variable selection using SVM-based criteria. *J Mach Learn Res.* 2003;3:1357-1370.
32. Díaz-Uriarte R, De Andres SA,. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics.* 2006;7(1):1–13.
33. Xia S, Yang Y. An iterative model-free feature screening procedure: forward recursive selection. *Knowl Based Syst.* 2022;246:108745.
34. Kursa MB, Rudnicki WR,. Feature selection with the Boruta package. *J Stat Softw.* 2010;36:1-13.
35. Keany E. BorutaShap 1.0.16. 2021. Accessed April 24, 2023.
36. Tibshirani R. Regression shrinkage and selection via the lasso. *JJ R Stat Soc Ser B Stat Methodol.* 1996;58(1):267-288.
37. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol.* 2005;67(2):301-320.
38. Scornet E. *Trees, Forests, and Impurity-Based Variable Importance in Regression*. Institut Henri Poincaré; 2023:21-52.
39. Nicodemus KK,. On the stability and ranking of predictors from random forest variable importance measures. *Brief Bioinform.* 2011;12(4):369-373.
40. Rudnicki WR, Kierczak M, Koronacki J, Komorowski J. *A statistical method for determining importance of variables in an information system*. Springer; 2006:557-566.
41. Grömping U. Variable importance assessment in regression: linear regression versus random forest. *Am Stat.* 2009;63(4):308-319.
42. Neville P. Controversy of variable importance in random forests. *JUST.* 2013;1(1):15-20.
43. Debeer D, Strobl C. Conditional permutation importance revisited. *BMC Bioinformatics.* 2020;21(1):1-30.
44. Rotari M. Conditional-Boruta. GitHub repository; 2021.
45. Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. *Ann Statist.* 2015;43(5):2055-2085. doi: https://doi.org/10.1214/15-AOS1337
46. Candès EJ, Fan Y, Janson L, Lv J. Panning for gold: 'model-X' knockoffs for high-dimensional controlled variable selection. *J R Stat Soc Ser B Stat Methodol*. 2018;80(3):551-577.
47. Genuer R, Poggi JM, Tuleau-Malot C. VSURF: an R package for variable selection using random forests. *R J.* 2015;7(2):19-33.
48. Yeh HP, Meinert K, Bayat M, Hattel J. Part-scale Thermo-mechanical Modelling for The Transfusion Module in The Selective Thermoplastic Electrophotographic Process. Volume 1000 Manufacturing and Materials Processing; 2022.
49. Stichel T, Geißler B, Jander J, Laumer T, Frick T, Roth S. Electrophotographic multi-material powder deposition for additive manufacturing. *J Laser Appl.* 2018;30(3):032306.

How to cite this article: Rotari M, Kulahci M. Variable selection wrapper in presence of correlated input variables for random forest models. *Qual Reliab Eng Int*. 2023;1-16. https://doi.org/10.1002/qre.3398

AUTHOR BIOGRAPHIES

Marta Rotari is currently a PhD Candidate in the Department of Applied Mathematics and Computer Science, Technical University of Denmark. She received her Master degree in Applied Mathematics from the University of Udine, Italy and her Bachelor degree in Mathematics from the University of Parma, Italy. Her research interest include statistical modeling, supervised and un-supervised learning methods and data analytics.

Murat Kulahci is a Professor in the Department of Applied Mathematics and Computer Science at the Technical University of Denmark and a Professor in the Department of Business Administration, Technology, and Social Sciences at Luleå University of Technology in Sweden. His research focuses on design of physical and computer experiments, statistical process monitoring, time series analysis and forecasting, and financial engineering.

CHAPTER 4

Causality and Correlation

4.1 Introduction

In the previous chapter, the focus was predominantly on process understanding, which, together with process improvement and optimization, constitutes the most significant aims discussed in Section 2.4. Machine Learning models facilitate the identification of correlations, or more broadly, relationships between the input and output data, often using observational data. For optimization, it is necessary to determine the input variables or machine parameters and their respective optimal levels to ensure the attainment of high-quality standards for the final products. Conducting effective optimization studies necessitates establishing a causality path that links the input data to the output data; see Figure 4.1. This causality path enables a comprehensive understanding of how changes in the input variables influence the output variables.

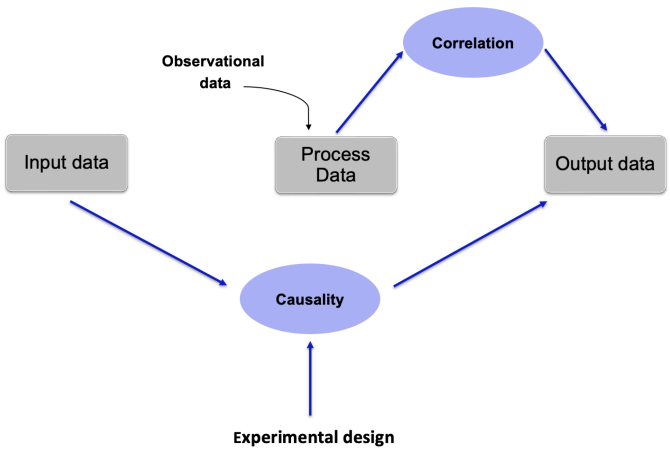


Figure 4.1. Correlation and Causality paths.

Establishing a causal relationship between input and output variables has significant relevance in many applications. Causality pertains to the association between a cause and its effect, where a modification in the input variable results in a corresponding alteration in the output variable. Specifically, suppose the variable X exhibits a causal effect on the variable Y. This implies that manipulating X while maintaining all other variables constant induces a change in the probability distribution of the Y variable. Causal inference enables us to anticipate the system's response to hypothetical scenarios or interventions by predicting outcomes for scenarios that were not directly observed. Causality grants us valuable insights into the underlying mechanisms and interdependencies within the system. Identifying causal relationships and optimal input levels that lead to the best outcomes leads to maximizing efficiency and performance.

Experimental design is commonly employed to establish causation. Randomised controlled experiments are frequently considered the most reliable method for establishing causal relationships. In experimental design, one or more independent variables (inputs) are manipulated while controlling all other variables. The dependent variable (output) is then measured to observe the effect of the independent variables.

The new 3D printing technologies are more complex production processes. This complexity lead to a significant increase in the number of initial parameters involved in the production process. As shown in Figure 4.2, a multitude of input parameters and sub-parameters are present, each with integer or real values. In this case, conducting extensive experiments on all the input variables is unfeasible due to the necessity of a substantial number of experiments, which would require a significant allocation of time and resources.

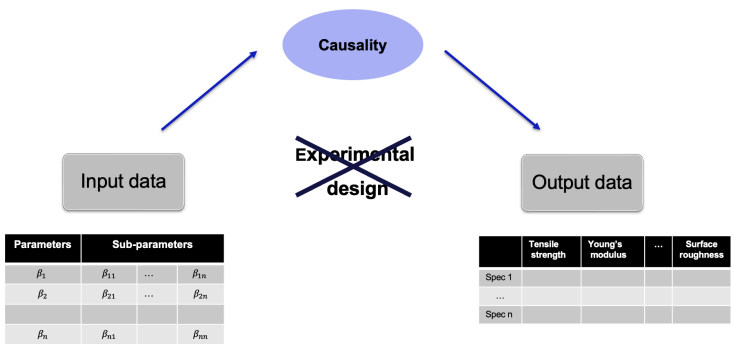


Figure 4.2. Input data include parameters and sub-parameters. No experimental design is possible in this context.

In situations when experimental procedures are not feasible, it becomes essential to incorporate a preliminary stage that enables the reduction of input variables or offers

direction for selecting suitable inputs to begin the exploration process. The selection of these variables is frequently reliant upon technical knowledge and an understanding of physical principles. Unfortunately, in this scenario, where these technologies are still largely unknown, there is the need to carry out data-driven exploratory analysis.

To tackle this challenge, we have adopted a novel approach which introduces a preventive phase. The main objective of this initial step is to establish boundaries inside the input space through the use of observational data. In our particular situation, we will be employing process data to analyse its correlation with the output. This analysis aims to identify probable causal variables that may influence the output. Through the analysis of the correlation between input variables and the resultant output, significant insights may be obtained into the underlying linkages that govern the behaviour of the system. While it is important to note that correlation does not necessarily imply causation, it plays a crucial role in initiating the exploration of the relationship between variables and can assist in identifying potential areas for further investigation. Based on the insights acquired from this preliminary step, the resources can be allocated and concentrated adequately. This will greatly improve the efficiency and efficacy of the entire causal research. The method, along with an applied example, is presented in the following paper titled ”Correlation to Causality”.

4.2 Paper: From correlation to causality

Marta Rotari and Murat Kulahci, *”From correlation to causality”*, to be submitted at Quality Engineering Journal.

From correlation to causality

Murat Kulahci[a,b], Marta Rotari[a]

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

Keywords: Correlation; causality; causal relations; observational data; designed experiments; large input space

Abstract

Causal models and experimental design are commonly used methodologies for establishing causal relationships. However, the use of such methods might pose challenges in several scenarios, particularly when examining novel and unknown systems, rendering the task almost unattainable. In scenarios where conducting experimental design is infeasible due to a large number of input parameters, observational data may be utilized as an initial basis for uncovering connections and acquiring insights to delineate the input space. In this study, we propose a hybrid approach that combines the analysis of correlation using observational data and machine learning techniques, followed by the utilization of designed experiments to establish causality. We demonstrate a practical application of our approach through a case study in the field of additive manufacturing.

1. Introduction

Causality plays a crucial role in numerous areas, such as science, medicine, production and many others. Understanding the cause-effect relationship between variables is essential for comprehending phenomena, predicting outcomes, optimizing the processes and making informed decisions. Causation refers to a relationship in which one event (the cause) immediately causes or influences the occurrence of another event (the effect). It suggests that changes in the cause result in predictable changes in the effect, indicating a cause-and-effect relationship. Controlled experiments, rigorous observational studies and causal models are usually used to establish causality. However, determining causality remains a very complex task that has often been plagued by the confusion between causality and correlation.

In statistical analysis, correlation is a statistical measure that quantifies the degree of association or relationship between two or more variables based on observational data. Observational data refers to information collected from one or multiple variables of interest, records without intervention or manipulation by the researcher. These variables may represent characteristics, attributes, or behaviours being studied. The correlation provides an estimate of the strength and direction of the relationship between the variables. Although a strong correlation between two variables suggests a relationship, correlation does not necessarily imply a causal relationship [1, 2, 3]. Additional analysis and evidence are required to establish a causal link between variables [4, 5]. Other variables, referred to as confounding variables, may affect the observed relationship.

While correlation serves as a valuable measure for understanding the relationship between variables, it is important to recognize that inferring causality requires a more rigorous approach, typically involving experimental design. Experimental design entails deliberately manipulating variables and observing the resulting effects to establish a cause-and-effect relationship. However, conducting experiments to determine causality in numerous situations may not be feasible or practical. Causal models, such as Directed Acyclic Graphs (DAGs) or Structural Equation Modelling (SEM), are employed as a means to deduce causation. These models are methodologically robust and have widespread adoption across several disciplines. Nevertheless, their use may be limited by significant complexity, time constraints, or other practical limitations.

While experimental data and causal models are often considered to be expensive and time-consuming, using existing data, often observational data, and utilising correlation analysis as a first step might provide a useful indication

in the pursuit of causation. When experimental designs are unavailable, correlation analysis can be used as a preliminary approach, providing clues and insights into potential causal relationships. While correlation does not establish causality definitively, it can highlight associations and guide researchers towards further investigation. This study addresses the issue of causality when direct experiments are not feasible. Our proposed approach involves utilizing a preliminary phase on the observational data to be used as an indicator for initiating the search for causality.

To illustrate the application of our approach, we considered a case study in the field of Additive Manufacturing (AM). Additive manufacturing, often referred to as 3D printing, has garnered significant attention in recent years due to its potential to revolutionize manufacturing processes. By examining the process variables and variables selection technique, it is possible to identify a subset of initial parameters to drive experimental designs related to causality determination.

The following section is devoted to the concepts of correlation and causality, exploring how causality has been historically perceived and how modern models approach the notion of causality. Section 3 describe the limitation of causal model and experimental designs. Inferring causation from observational and historical data is described in Section 4. Subsequently, in Section 5, we describe the steps of a possible hybrid approach to determine causality. This approach is then applied to a real case of additive manufacturing in Section 6. Finally, some considerations and conclusions.

2. Causality

Before the early 20th century, the distinction between correlation and causality was not as well understood and defined as it is nowadays. Although scientists have acknowledged the notion of two variables being related or varying together, the idea of one variable causing changes in another was not always clear. During this period, the discipline of statistics was in its early developmental phase, and researchers frequently relied on basic correlations to infer associations between variables. Observational data was frequently utilised to establish conclusions and researchers frequently failed to include confounding variables. These variables are not the main focus of a study but can affect the relationship between the dependent and independent variables. They are often associated with both the dependent and independent variables, and their presence can lead to misleading or incorrect conclusions about the true relationship between those variables.

For example, consider a scenario where there is a positive correlation between ice cream sales and drowning deaths. In the absence of a comprehensive knowledge of causation, it might be concluded that buying more ice cream increases the likelihood of drowning. In reality, both variables are subject to the impact of a third component, namely, the temperature. Elevated temperatures are associated with augmented ice cream purchases, as well as a rise in individuals engaging in swimming activities, hence amplifying the potential hazards of drowning. In this particular scenario, the temperature serves as a confounding variable that elucidates the observed association between ice cream sales and drowning fatalities.

As research advanced and the discipline of statistics faced further development, scientists and statisticians began to realize the importance of distinguishing correlation from causation. They recognized that establishing causality requires more rigorous methods, including experimentation and the development of causal models. In the early and mid-20th century, the distinction between causality and correlation was established, and this signed the big transformation in social science, statistics, medicine and medical trials. Sir Ronald A. Fisher, a prominent statistician and geneticist, significantly contributed to this field.

Sir Ronald A. Fisher's work can be traced back to his involvement in agricultural research during the 1920s and 1930s, where he developed experimental designs to assess the causal relationships between various agricultural factors and crop yields. He made significant contributions to the fields of statistics and experimental design, including concepts related to correlation and causation. Correlation, as described by Fisher, represents a statistical measure that quantifies the degree of association between two or more variables. Causation, on the other hand, involves demonstrating a cause-and-effect relationship between variables. Fisher emphasized the importance of experimental design in establishing causation. He advocated for randomized controlled experiments, where participants are randomly assigned to treatment and control groups, allowing for the isolation of the effect of the independent variable on the dependent variable. Fisher's work laid the foundation for experimental design and statistical inference. He explored concepts like confounding, randomization, and the importance of controlling for variables to establish causal relationships in observational studies [2, 6, 7].

During the same period, the early and mid-20th century, there was a notable emergence of core concepts in the field of causal modelling [8]. The domain of causal inference and causality modelling was a significant expansion, leading to the development of a new class of causal models. The origins of these models may be traced back to the disciplines of statistics, econometrics, and social sciences.

In 1921, the researcher Sewall Wright introduced a method called ”path analysis”. This work later evolved into the Directed Acyclic Graphs (DAGs). Judea Pearl made substantial contributions to the progress and comprehension of DAGs in the field of causal inference [9, 1]. The Directed Acyclic Graphs are a valuable instrument for comprehending and visualizing the complex causal pathways between variables [10]. DAGs can be used to help establish causality by identifying potential causal pathways and excluding alternative explanations. It can be used to identify potential sources of bias or confounding and determine which variable may be necessary to control for in order to establish a causal relationship between two variables. DAG is also a visualization tool. The variables are represented as nodes and the causal relationships between variables are represented as directed edges or arrows. The direction of causality is indicated by the arrows, with the arrow's tail representing the cause and the arrow's head representing the effect. DAGs permit the representation of confounding variables.

DAGs offer a transparent and interpretable means to visualize complex causal relationships among variables, and it is a valuable tool that helps to avoid causal inference pitfalls. However, DAGs alone do not establish causality and require complementary methods as experimental designs. Constructing DAGs could be challenging, demanding a clear understanding of underlying causal relationships. They are models sensitive to the strong assumptions and the quality of the data. DAGs facilitate comprehension of causal relationships but lack standalone causal establishment capability.

Around the 1970 was introduced the concepts of latent variables and covariance structure models. This framework forms the basis for the Structural Equation Modeling (SEM) [11, 12]. Structural Equation Modeling is a statistical technique employed to test and validate complex relationships between variables and estimate the magnitude and direction of these relationships while also identifying the causal mechanisms that underpin them. In SEM, the variables are represented as either latent variables or observed variables, and their relationships are represented through a set of equations that identify direct or indirect effects. In SEM, causality is used to establish the direction of the relationships between variables and is represented in the model by directional arrows. Causality is a crucial concept in SEM, but establishing causal relationships requires additional techniques beyond SEM analysis, such as experimental designs. The model can only provide evidence for or against causal relationships using the available data. Although SEM is a powerful tool for examining complex relationships between variables, it is crucial to recognise its inherent limits. The model requires a large sample size in order to accurately estimate the parameters of the model and assess its sensitivity. It is also computationally complex and time-consuming. In addition, interpreting SEM results can be difficult due to the numerous test statistics and parameters. Even though SEM can test for causal relationships, it cannot establish causality on its own.

In the second half of the 20th century, it has been introduced the Instrumental variables (IV), which is a statistical technique widely utilized to determine causal effects in causal inference and machine learning [13]. It is primarily used when dealing with confounding variables. To control for unmeasured confounding, a third variable is introduced, named instrumental variable (IV), which is a cause of input variables and has no direct effect on the outcome. The instrumental variable is used to identify the true correlation between the explanatory variable and the outcome. Two well-known models for estimating treatment effects using an instrumental variable are the two-stage least squares estimator (2SLS) and the control function estimator (CFN). In the context of machine learning, instrumental variables are used to address confounding bias in observational studies or experiments with non-randomized treatments for data mining and explanatory analysis purposes. However, there are limitations to the use of instrumental variables, such as the requirement of the strong assumption that there is no direct relationship between the instrumental variable and the outcome. In addition, the instrumental variable must be independent of any unmeasured confounding variables that may influence the outcome. This requirement is hardly and rarely satisfied in practice.

3. Limitations of causal models and Experimental Designs

Causal models, such as Directed Acyclic Graphs and Structural Equation Modelling, are widely recognised as robust methodologies for comprehending and examining causal relationships. Nevertheless, applying these models might be challenging or less feasible in some circumstances and domains, such as additive manufacturing.

The construction of causal models as DAGs or SEM requires a clear understanding of the process or the system, including potential confounding factors and intermediate variables. In situations where the systems are extremely complicated, constructing accurate DAGs or SEM models becomes a challenging problem. Building reliable causal models becomes extremely problematic if the underlying processes are not well-defined or fully understood. High-dimensional problems prevent the construction of causal models. Indeed, constructing a comprehensive DAG or SEM model that considers many variables and their interactions can be computationally intensive and prone to model complexity issues.

For example, additive manufacturing processes involve multiple intricate and complicated steps, leading to complex causal relationships and can involve many confounding variables. These novel production systems lack widespread recognition and are still under investigation processes. The processes often involve complex sub-processes and interactions that may lead to many confounding variables that have not yet been identified. The manufacturing processes are still undergoing development and optimisation, and as a consequence, they could exhibit variability due to variables such as machine calibration and environmental conditions. This variability can make it difficult to establish clear cause-and-effect relationships, as different runs might yield different results. Moreover, additive manufacturing often involves many variables, such as material properties, different temperatures, pressure, printing parameters, environmental conditions and post-processing steps. Building a model considering all these variables and their interactions poses challenges in establishing a causal model.

A further factor that limits the use of causal models is the existence of emergent phenomena within many systems, where unexpected interactions or properties emerge from the complex interplay of variables. Traditional causal models designed for more linear systems might not easily capture these phenomena. Systems that exhibit nonlinear behaviours, where small changes in input variables can lead to significant and non-proportional changes in output. Traditional causal models might struggle to capture these nonlinear relationships accurately. Validating causal models is crucial to ensure their accuracy and reliability. Ultimately, the construction of such models frequently necessitates a substantial investment of both time and resources. In certain instances, particularly in rapidly evolving industries like additive manufacturing, time constraints might limit the thorough construction and validation of complex causal models.

Various constraints frequently hinder the use of causal models in certain circumstances and fields. Frequently, the same circumstances provide significant challenges for conducting experiments. In the context of additive manufacturing technologies, the manufacturing processes are extremely innovative and provide extensive customization capabilities. The process in question is intricate and has several sequential stages. This phenomenon involves having a wide variety of input parameters that must be established at the beginning of an order. The optimization of these processes involves establishing and determining the level of each parameter to ensure the high quality of the manufactured products. Due to the extensive number of input variables and their potential combinations, conducting extensive experimentation becomes unfeasible.

Designing and executing experiments can be resource-intensive, requiring substantial financial investment. The costs can escalate when multiple experiments are necessary to investigate the causal relationship between variables thoroughly. Experiments often require a significant amount of time to collect data, especially when studying long-term effects or phenomena that evolve over time. Due to these factors, it frequently becomes impracticable to carry out extensive experiments, necessitating the utilisation of alternative approaches.

4. Causation using observational and historical data

While randomized controlled experiments are considered the gold standard for establishing causation, they are not always feasible. Assessing causation from observational data can be challenging but is possible under certain circumstances. In the absence of experimental control, various strategies could be employed to gather evidence for causation from observational data.

One such strategy is the utilization of multiple sources of evidence. It is possible to triangulate information from different sources to strengthen causal claims. Drawing upon diverse archival records and other primary sources allows for a more comprehensive understanding of the historical context and potential causal mechanisms.

Another strategy involves examining temporal precedence. Demonstrating that the cause precedes the effect in historical data can provide support for causal arguments. Analyzing the timing and sequencing of events can offer

valuable insights into potential causal relationships. Consistency of patterns across multiple historical cases or contexts is another crucial aspect in establishing causation. Identifying consistent patterns across multiple historical cases or contexts can provide further evidence for causation. If similar causal relationships are consistently observed, it strengthens the plausibility of a causal connection. Identifying and highlighting these consistent patterns can further strengthen the evidence for causation.

It is important to acknowledge the limitations of assessing causation from historical data. Historical observational data often involve incomplete or biased records, challenges establishing causality due to limited data availability, and the potential influence of unobserved or unmeasured variables. These limitations highlight the need for rigorous analysis, careful interpretation, and critical consideration of alternative explanations in historical causation research.

In situations where the input space is extensive, the use of observational data is necessary to narrow the input space. In such situations, Partial Least Squares (PLS) regression-based techniques provide an appealing solution by analyzing the observational data in conjunction with corresponding output data. These methods facilitate the inference of causality within a reduced space, and several approaches based on Latent Variable Regression Model Inversion (LVRMI) have been proposed in the literature [14, 15, 16]. These methods aim to define the Null Space (NS), which represents the projection onto the latent space of all possible combinations of inputs that theoretically result in the desired outputs. However, the existing methods lack a unified approach to define the Null Space and are confined to the domain of historical products that have already been developed. The analytical expression of the NS exists under the assumption that at least one combination of the input space exists. Consequently, there is currently no standardized procedure to define the NS.

5. Proposed approach

In order to establish a causal relationship between the input and output domains, it is imperative to comprehend the causal mechanisms through which certain input variables influence the result. This comprehension may be achieved by employing a causal model and conducting experiments that establish the relationships between input and output variables. However, in situations when the system is novel, limited knowledge is available and there is a lack of prior information, the implementation of a causal model becomes very challenging. Additionally, running experiments may be prohibitively expensive or need enormous resources. Indeed, in several instances where the relationship between variables is under investigation, we are faced with many input variables to examine without any clear guidance on where to start the exploration.

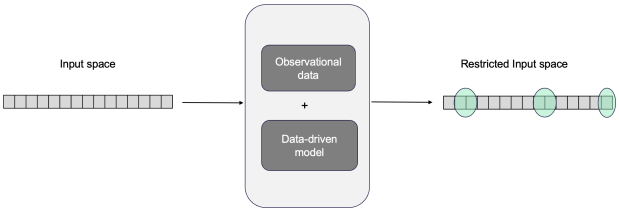


Figure 1: Narrowing the input space through the observational data and data-driven models.

In the case of a large size of the input space, conducting exhaustive experimental designs covering all input parameters and all possible combinations becomes impractical and infeasible. To overcome this limitation, a possible solution is to identify a subset of input variables to start exploratory experiments. To pursue this objective, a viable solution is the utilization of data-driven models built based on observational data. For example, in the additive manufacturing case, the observational data are the process data, information collected during the production process. When

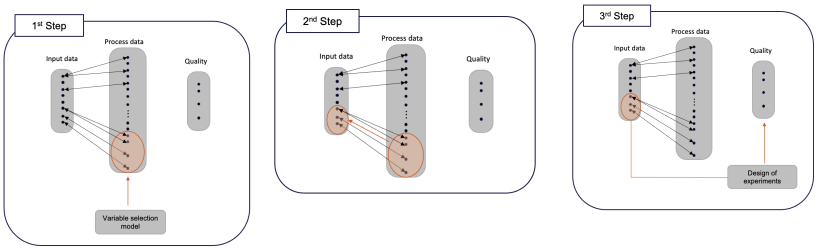
conducting numerous experiments would be prohibitively expensive or time-consuming, an initial investigation using observational data can help narrow the focus on a subset of the input space and guide subsequent experimental efforts, Figure 1.

The analysis of observational data serves as a valuable initial step in exploring the relationship between variables and response. To this end, employing supervised machine learning models, such as regression models or tree-based models, offers significant advantages by providing estimations of individual variable contributions towards quality prediction. Additionally, when dealing with a large number of variables, utilizing models with variable selection capabilities becomes advantageous. This problem, commonly known as the variable selection problem, aims to identify the most influential variables that strongly affect quality.

The variables selection models yield a subset of relevant process variables that are deemed meaningful from a predictive standpoint, see Figure 2a. These variables provide valuable insights and serve as preliminary indicators, directing a further investigation into the specific aspects of the production process parameters that should be explored in greater depth at a later stage.

The analysis of variable selection results offers the opportunity to identify and highlight process variables that exhibit promising relationships with quality, thereby justifying the need for further investigation. Drawing upon the expertise of domain specialists, it becomes feasible to pinpoint relevant process parameters within the input space that are associated with the selected process variables, as depicted in Figure 2b. These identified input variables constitute a subset of the overall input space and become the primary focus of subsequent experimental investigations. Through controlled experiments, the causal relationship between these variables and output quality can be further explored. The deliberate selection of a subset of variables for experimentation effectively reduces the complexity and magnitude of the input space, enabling targeted experiments to be conducted within a manageable scope.

Consequently, by focusing on the selected variables derived from this connection with the process data, researchers can strategically design and execute experiments aimed at directly investigating causal relationships. This targeted approach enables the efficient allocation of resources and reduces the overall experimental space by narrowing down the factors that require investigation. Moreover, these selected variables serve as prime candidates for experimental studies, as they are more likely to exhibit causal relationships, Figure 2c.



(a) Variable selection on the process data. (b) Connection with the input space. (c) Design of Experiments on the Reduced Input Space.

The combination of data-driven analysis using process data and subsequent experimental studies provides a pragmatic and efficient way to navigate the correlation versus causation challenge. While the initial correlation analysis helps guide the selection of variables for experimentation, the subsequent experimental studies enable the establishment of causal relationships with more confidence and rigour. This approach optimizes resources and allows researchers to gain valuable insights into causation while maximizing available data and minimizing costs.

6. Application

This section provides a real-world application of the proposed approach in Additive Manufacturing Production. Additive manufacturing (AM), also known as 3D printing, is an innovative manufacturing process that enables the production of products with complex geometries and high levels of precision and efficiency. It is a revolutionary

approach to manufacturing in which objects are constructed layer by layer using digital designs. While additive manufacturing has demonstrated great promise, it is still largely unexplored and faces numerous challenges that require optimization. Many studies are conducted aimed at process understanding and process optimization [17, 18]. One of the key problems is analyzing process parameters and identifying the input parameters and their optimal level that ensure high and uniform quality standards.

In additive manufacturing and, in general, all production processes, a comprehensive understanding of the causal relationship between input parameters and output quality plays a crucial role in achieving process optimization and control. Multiple data sources can be identified within the AM process, which can be categorized as follows. The input space encompasses machine settings or process parameters that are established at the beginning of an order. Throughout the production chain, diverse sensors are strategically placed to gather data during production, constituting the process data that encompasses a wide range of production-related information. The output data reflects the quality of the manufactured products, such as mechanical properties (Tensile strength and Young's Modulus) or other defects such as dimensional defects, roughness Etc. The primary objective is to identify the input parameters that most impact the quality, thereby establishing a causal relationship between the input and output spaces.

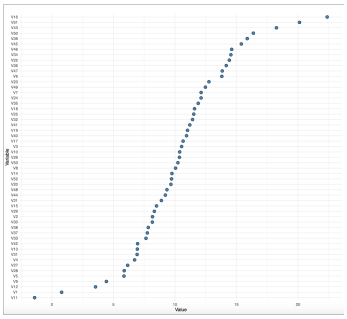


Figure 3: Ranking of the variables.

Emerging technologies are characterised by a greater complexity to their control systems, resulting in a large number of input parameters that must be configured at the outset of the process. The large number of input parameters, render it impractical to conduct an experimental design that covers the entire parameter range and their combinations. To address this challenge, an alternative approach is adopted, as elaborated in Section 5.

As for the process data, in this analysis, we considered the data of 18 different batches. During the production process, 53 continuous variables were assessed, labelled as $V1$ to $V53$, for the sake of confidentiality. Our main objective was to investigate the mechanical properties of the final products, specifically the tensile strength, which represents the maximum tensile stress a material can endure before breaking or deforming. Upon analyzing the data, we observed that the variables under study exhibited significant correlations. Initially, we employed the Random Forest model to establish the relationship between input and output variables. Subsequently, we utilized the mean decrease in accuracy measure to rank the variables based on their importance in the prediction process (see Figure 1). However, relying solely on the ranking made it challenging to determine the precise number and identity of variables that genuinely influence the output. Furthermore, due to the presence of high correlation among the input variables, the ranking's reliability was compromised, as the strong correlations led to an overestimation of variable importance. To address this issue and select the most relevant variables, we employed the Conditional Boruta model [19], specifically designed to handle correlated variables.

The outcome of the variable selection models is depicted in Figure 4. Notably, three variables have emerged as relevant: $V36$, $V49$, and $V39$. Each of these variables contributes to specific aspects of the manufacturing process. Firstly, variable $V36$ is associated with the creation of each micro-layer. This variable holds relevance in under-

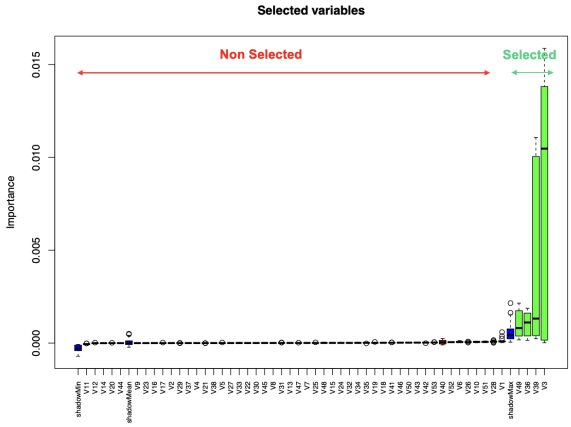


Figure 4: Variable selection results.

standing the fundamental building blocks of the manufacturing process. Secondly, variable $V49$ represents the fusion temperature of each layer. Temperature plays a pivotal role in determining the cohesion between molecules, making it a crucial factor influencing the physical characteristics of the final product. Lastly, variable $V39$ pertains to light exposure, which gradually deteriorates over time.

The selected variables by the Conditional Boruta model were subsequently linked to the input variables by the expert engineers. The input variables thus identified are denoted as Factors $1, 2$ and 3. By establishing this connection, a focused subset of the input data was identified in order to initiate the exploratory investigation. Subsequently, the third step of our approach, as depicted in Figure 2c, involves designed experiments. This experiment phase aimed to investigate the relationship between the aforementioned three factors and the desired outcome, the tensile strength. To accomplish this, we conducted a 2^3 design consisting of eight experimental runs without replications.

A full model with main factors and their combinations was fitted. The Normal plot related to the outcome is displayed in Figure 5. The analysis of the Normal plot revealed evidence regarding the importance of two main factors, 2 and 3. A model with only these main effects was fitted and the resulting coefficients with the related p-values are summarized in Table 1. The response surface plot, in Figure 5, illustrates the relationship between the identified factors and the resulting tensile strength. It demonstrates how variations in the levels of the factors influence the response variable. By examining this plot and considering the calculated coefficients, subsequent steps can be generated to further investigate and improve the quality.

Term	Estimate	Std Error	t Ratio	Prob > $\lvert t \rvert$
Intercept	37.800893	0.105485	358.35	¡0.0001
Factor2	0.5026786	0.105485	4.77	0.0050
Factor3	0.3151786	0.105485	2.99	0.0305

Table 1: Results

In summary, our study involved 18 batches of products, with tensile strength as the outcome variable. The presence of high correlation among the variables necessitated the use of a model specifically designed to handle such

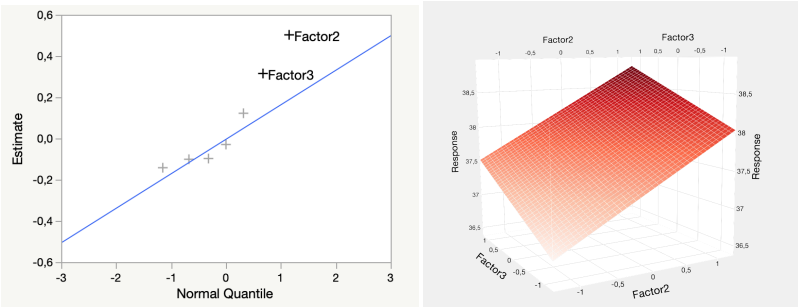


Figure 5: Normal plot and response surface.

correlation. The Conditional Boruta model selected three variables as relevant, which were then connected to the input variables. Three input variables were chosen for an experimental design, leading to eight runs. The experiment's outcome revealed that two of these factors significantly influenced tensile strength, highlighting their importance in determining the quality of the final products.

7. Conclusion

In numerous scenarios, establishing a comprehensive understanding of the causal relationship between input parameters and output quality is essential for multiple purposes, such as optimization or understanding. However, conducting an experimental design to assess causality becomes impracticable when faced with an extensive input space. To overcome this limitation, an alternative approach emerges, whereby observational data is involved in gaining insights into a specific region of the input space suitable for initiating exploratory experiments. By employing data-driven models, it becomes feasible to establish connections between the observational data and output data, facilitating the identification of the most relevant variables. The link between the relevant variables and the input parameters identifies a starting point for subsequent experimentation. It enables the identification of a representative subset of the input space that can be subjected to further experimental investigations, delving deeper into the causal relationship between the input parameters and output data. This approach is illustrated in the case study in Additive Manufacturing. Our findings led to the identification of two significant factors that exerted a substantial influence on the considered quality outcome. These results underscore the proposed methodology's value and ability to guide the discovery of influential factors in complex manufacturing processes.

References

[1] J. Pearl, Causality, Cambridge university press, 2009.
[2] R. Fisher, Statistical methods for research workers, 1st edn edinburgh (1925).
[3] K. Pearson, The grammar of science, 3rd, London, England: A. and C. Black 153 (1911).
[4] J. de Mast, S. H. Steiner, W. P. Nuijten, D. Kapitan, Analytical problem solving based on causal, correlational and deductive models, The American Statistician 77 (1) (2023) 51–61.
[5] M. A. Hernán, J. Hsu, B. Healy, A second chance to get causal inference right: a classification of data science tasks, Chance 32 (1) (2019) 42–49.
[6] R. Fisher, The design of experiments ((1935, 1st), Edinburgh: Oliver and Boyde (1960).
[7] R. A. Fisher, Statistical methods and scientific inference. (1956).
[8] N. Barrowman, Correlation, causation, and confusion, The New Atlantis (2014) 23–44.
[9] J. Pearl, Causal inference in statistics: An overview (2009).
[10] A. P. Dawid, Beware of the dag!, in: Causality: objectives and assessment, PMLR, 2010, pp. 59–86.
[11] R. Weston, P. A. Gore Jr, A brief guide to structural equation modeling, The counseling psychologist 34 (5) (2006) 719–751.
[12] K. A. Bollen, Structural equations with latent variables, Vol. 210, John Wiley & Sons, 1989.

[13] A. Wu, K. Kuang, R. Xiong, F. Wu, Instrumental variables in causal inference and machine learning: A survey, arXiv preprint arXiv:2212.05778 (2022).
[14] D. Palací-López, P. Facco, M. Barolo, A. Ferrer, New tools for the design and manufacturing of new products based on latent variable model inversion, Chemometrics and Intelligent Laboratory Systems 194 (2019) 103848.
[15] Z. Liu, M.-J. Bruwer, J. F. MacGregor, S. S. Rathore, D. E. Reed, M. J. Champagne, Modeling and optimization of a tablet manufacturing line, Journal of Pharmaceutical Innovation 6 (2011) 170–180.
[16] J. MacGregor, M. Bruwer, I. Miletic, M. Cardin, Z. Liu, Latent variable models and big data in the process industries, IFAC-PapersOnLine 48 (8) (2015) 520–524.
[17] T. Stichel, B. Geißler, J. Jander, T. Laumer, T. Frick, S. Roth, Electrophotographic multi-material powder deposition for additive manufacturing, Journal of Laser Applications 30 (3) (2018) 032306.
[18] H.-P. Yeh, K.Meinert, M.Bayat, J.Hattel, Part-scale thermo-mechanical modelling for the transfusion module in the selective thermoplastic electrophotographic process, in: WCCM-APCOM 2022, Volume 1000 Manufacturing and Materials Processing, 2022.
[19] M. Rotari, M. Kulahci, Variable selection wrapper in presence of correlated input variables for random forest models, Quality and Reliability Engineering International (2023).

CHAPTER 5

Multi-group Analysis

5.1 Introduction

Process data is a significant component of the data pertaining to 3D printers. This Chapter explores the process data, its analysis and potential applications, emphasizing its crucial role in fully comprehending the manufacturing process.

Process data refers to the data acquired during manufacturing operations through strategically placed sensors integrated into the machinery. This data includes various aspects of the process such as machine performance, temperature levels at various stages, energy consumption, material utilization and other critical parameters. The process data offers a valuable resource for gaining insights into the process. First, it allows for a deeper understanding of the dynamics that govern the production process. Patterns, correlations, and cause-and-effect relationships could be identified that contribute to the overall process behaviour. It can be identified areas for improvement and devise strategies to optimize efficiency, reduce waste and enhance product quality. Hence, this data plays a crucial role in enhancing our knowledge of the production processes and supporting decision-making for process improvement and optimization.

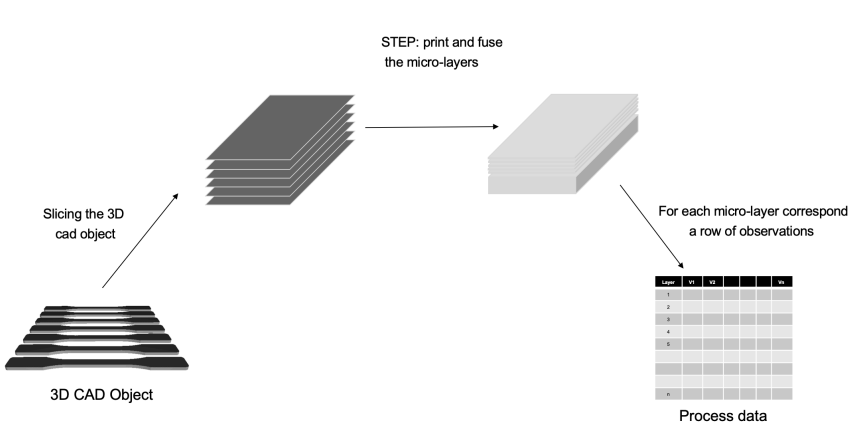


Figure 5.1. Data process acquisition.

In each individual production executed by the machine, a broad array of data is collected. Data is gathered pertaining to a total of 53 variables. An observation of each variable is saved for every micro-layer produced and melted as depicted in Figure 5.1. In order to effectively organize and structure the given data, a matrix or two-way array structure is utilized. In this structure, the variables are represented by columns, while rows represent the observations. This configuration enables a methodical depiction of the gathered data, hence allowing later analysis and interpretation. This data allows for systematic analysis and exploration of the relationship between variables and layers. It can be used for online monitoring and facilitates a comprehensive understanding of the production process and machine performance.

The development of these novel AM technologies has been specifically aimed at facilitating large-scale production. Consequently, the machines are capable of producing many goods in a single batch, arranged in ordered rows. Production can be conducted in small batches, either as a single row or as a large batch consisting of two or more rows. In Figure 5.2, three rows of products are illustrated. Nevertheless, it is important to acknowledge that the machine's complete production capability enables the possibility of manufacturing a far greater quantity of product rows. This can vary based on factors such as machine capabilities, production requirements and operational considerations.

Figure 5.2. Process data and products.

The manufacture of more than two rows of products follows the same production principle, with only a few changes. The production takes place according to many micro-layers, which are pressed and fused together to create the final products, as described in Section 2.3.2. Additional support layers are placed between each row of products to provide structural integrity and facilitate the division between the individual products. These support layers serve as temporary structures and are later dissolved through a chemical wash after the completion of the production process.

Since each micro-layer corresponds to a row of observations in the process data, it is possible to establish a precise connection between each row of products and its corresponding observations. Based on this characteristic, the observations acquired during the printing process can be categorized into distinct groups, see Figure 5.2.

The role and placement of the layers within the production process determine the grouping of the observations. This division allows for a more systematic analysis and understanding of the different stages of production. The following is an expanded description of the various groups of observations:

1. First Group of observations: **Support Layers for Base**
   The first group of observations corresponds to the few support layers that create a small support base and division between the building plate and the first row of products. These layers are specifically designed to provide a stable foundation for the subsequent layers and to separate the products from the building plate.

2. Second Group of observations: **Layers Forming the First Row of Products**
   The second observation group consists of layers forming the first row of products. These layers contribute directly to the structure and composition of the initial row of products in the 3D bulk.

3. Third Group of observations: **Support Layers Dividing Product Rows**
   Following the first row of products, the third group of observations corresponds to the support layers that divide the first and second rows of products. These support layers are strategically placed to separate the different rows and maintain their individual integrity during the production process. After production, these layers are dissolved in a chemical wash in order to facilitate the separation of all constituent products.

This grouping pattern continues for subsequent rows of products and support layers until the completion of the production process.

The process of categorizing the observations into separate groups promotes the analysis and interpretation of the data. This organization facilitates the examination of the effects of support layers, the production of individual product rows and highlights the evolution of the variables as the rows of products progress. Moreover, it fosters the opportunity to conduct focused analysis and enhance the performance of certain elements inside the manufacturing procedure.

In order to get an accurate understanding of the production process, it is necessary to take into account several iterations of batches. Even within the same production setup, small variations can occur. By analyzing multiple batches, a more accurate representation of the production mechanism can be obtained, taking into account the inherent variability that can arise. The different batches introduce a third dimension to the data array, transforming it into a three-way array, refer to Figure 5.3. Each batch corresponds to a distinct set of observations, capturing the same variables.

The three-way arrays thus obtained can be denoted as $\mathcal{X} \in \mathbb{R}^{N \times M \times K}$ whose elements are x_{nmk} where $n = 1, \ldots, N$, $m = 1, \ldots, M$ and $k = 1, \ldots, K$. In

a multi-group structure, the observations are divided into G groups. Each group $g \in \{1, \ldots, G\}$ is represented by $1, \ldots, n_g$ rows, such that $\sum_{g=1}^{G} n_g = N$. A visual representation of this structure is shown in Figure 5.3.

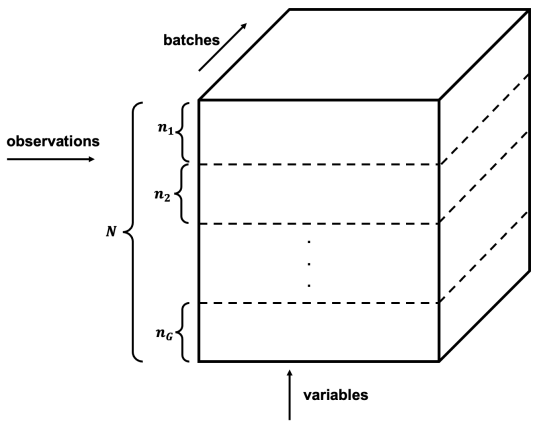


Figure 5.3. Three-way array with a multi-group structure.

The resulting three-way array encapsulates a multi-group structure. This multi-group structure enables a more comprehensive analysis and understanding of the production process, taking into account both within batch and groups and between batch and group variations. It allows for the identification of patterns, trends, and relationships within and across batches and groups, facilitating improved process control, quality assurance and optimization efforts.

In general, the data that present a multi-group structure are data sets where a set of variables is observed in different groups, each representing a specific aspect or characteristic. These groups could be formed based on various factors, such as demographic features, geographic locations, experimental conditions, or any other relevant categorical variables. An example of a multi-group two-way array is the Iris data set. It is a classic and well-known data commonly used in statistical analysis and machine learning. The Iris data set consists of measurements of iris flowers from three different species: setosa, versicolor, and virginica. Each species forms a group. The columns represent different variables measured for each flower, while the rows correspond to individual observations or samples of iris flowers. The four features recorded for each flower are the sepal length, sepal width, petal length, and petal width, all measured in centimetres.

In the context of research or analysis, the presence of a multi-group structure in the data is crucial, as it allows for the examination of how different variables behave in different groups. In Section 5.3, a new unsupervised methodology is presented,

which extends the PARAFAC algorithm to the case of multi-groups. The PARAFAC algorithm is a widely used decomposition technique for multi-way arrays; a brief overview is given in the next section. This extended methodology builds upon the traditional PARAFAC algorithm and adapts it to handle the complexities introduced by the multi-group structure in the data. This new proposed methodology allows us to effectively extract latent factors common to all the groups providing valuable insights and facilitating further analysis. The model aims to capture the inherent structure and relationships present within the multi-group three-way array, allowing for improved understanding and interpretation of the data.

5.2 Multi-way array decomposition models

Multi-way array or tensor decomposition has emerged as a highly used technique in the analysis of complex, multi-way data structures. These methods are tri-linear or multi-linear extensions of classical Principal Component Analysis PCA on two-way data. In a diverse range of fields, from image processing and signal analysis to neuroscience and chemometrics, the need to uncover hidden patterns and underlying structures has driven the development of these methods. The Tucker [25, 26] and PARAFAC [27] models stand as prominent approaches within the realm of tensor decomposition, each offering a distinct perspective on how to disentangle the complexity of multi-way data arrays.

Both models have been extensively applied in many fields when the data are presented as multi-way arrays. These models can provide the capability to extract significant insights, decrease dimensionality and investigate complex relationships. The PARAFAC and Tucker models provide useful frameworks for analysing multi-dimensional data and uncovering hidden factors in multi-modal data. These models give significant perspectives that facilitate a deeper understanding and interpretation of the complexities present in such data sets. In the following section, a brief overview of the Tucker and PARAFAC model is given. A three-way array $\mathcal{X} \in \mathbb{R}^{N \times M \times K}$ is considered.

5.2.1 Tucker

The Tucker model, also known as Tucker decomposition or Tucker factorization, is a multivariate statistical technique used for multi-way array decomposition [25, 26]. The Tucker model is designed to capture the underlying structure of multi-way data by representing it as a core array multiplied by factor matrices along each mode. A representation of the Tucker model can be seen in Figure 5.4.

Given a three-way array $\mathcal{X}$, the Tucker model decomposes this array into a core array $\mathcal{G} \in \mathbb{R}^{F_1 \times F_2 \times F_3}$ and factor matrices $\mathbf{A} \in \mathbb{R}^{F_1 \times N}$, $\mathbf{B} \in \mathbb{R}^{F_2 \times M}$ and $\mathbf{C} \in \mathbb{R}^{F_3 \times K}$. The model can be written as follow:

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} + \mathcal{E}$$

where $\mathcal{E}$ is the residual three-way array and $\times_i$ represents the mode-i product between an array and a factor matrix.

The goal of the Tucker model is to find the core array $\mathcal{G}$, along with the factor matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, that minimizes the difference between the original array $\mathcal{X}$ and the reconstructed array $\mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$. This can be formulated as an optimization problem, often estimated using Alternating Least Squares (ALS) [28] or through Higher-Order Singular Value Decomposition (HOSVD) [28].

The Tucker model is a frequently used model in multi-way array decomposition, providing a formal and statistical framework for the identification of underlying patterns within multi-dimensional data sets. By decomposing the array into a core array and factor matrices, the model unveils the latent factors responsible for the observed data patterns, enabling their interpretation and analysis. The model allows for a flexible representation of the original array. The flexibility of the Tucker model lies in its ability to the specification of the core array dimensions and the rank of the factor matrices. This flexibility makes the model a versatile tool for handling high-dimensional data, dimensional reduction and extraction of underlying structures of the data. The Tucker model finds applications in various fields, such as image analysis, natural language processing, neuroscience, social network analysis, and more.

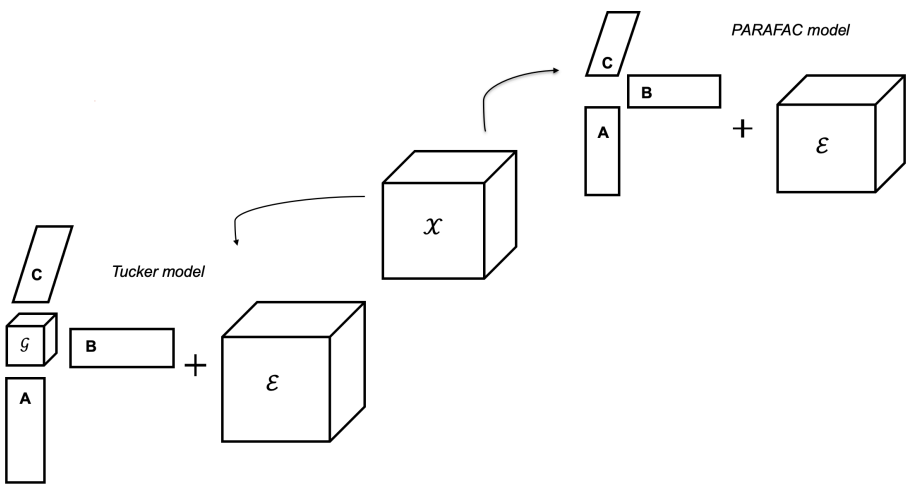


Figure 5.4. Representation of the Tucker and PARAFAC models.

5.2.2 PARAFAC

The PARAFAC (Parallel Factor Analysis) model is a multivariate statistical technique used for multi-way array decomposition and factor analysis [27]. It extends the concept of matrix factorization to higher-order data arrays. This model is designed to unravel the latent structure and underlying factors embedded within multidimensional data arrays. A representation of the PARAFAC model can be seen in Figure 5.4.

The PARAFAC model has numerous applications in various fields, such as chemometrics, neuroscience, psychometrics, image analysis, etc. It can be used for analyzing multi-sensor data, spectroscopic data, EEG data, and other forms of multi-way data where the relationships between dimensions are of interest.

The fundamental objective of the PARAFAC model is to decompose of multi-way data array in order to extract the underlying model loadings. The PARAFAC model seeks to estimate the original multi-way array by decomposing it into a summation of rank-one arrays. Each rank-one array represents a unique combination of factors from the corresponding mode. This decomposition encapsulates the latent structure of the data, allowing for a more parsimonious representation while retaining the essential features. The PARAFAC model is a powerful tool usually used for data compression, variable extraction and pattern recognition.

The PARAFAC model aims to approximate a given three-way array $\mathcal{X}$ by three loadings matrices $\mathbf{A} \in \mathbb{R}^{N \times F}$, $\mathbf{B} \in \mathbb{R}^{M \times F}$ and $\mathbf{C} \in \mathbb{R}^{K \times F}$. A PARAFAC model with F components can be written as follows:

$$\mathcal{X} = \sum_{f=1}^{F} \mathbf{A}^r \otimes \mathbf{B}^r \otimes \mathbf{C}^r + \mathcal{E} \tag{5.1}$$

where $\mathbf{A}^r$ represents the r-th column vector of the loadings matrix, $\mathcal{E}$ is the residual three-way array and $\otimes$ denotes the outer product operation.

The PARAFAC model aims to determine the loadings matrices that best approximate the original three-way array $\mathcal{X}$. The model aims to minimize the following criterion:

$$\min \sum_{ijk} \left(x_{ijk} - \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} \right)^2$$

The complexity of the decomposition is controlled by the rank F of the model. The most suitable number of components F is based on the available data and the specific objectives in question. If the Tucker3 core array is structured as a super-diagonal array, where the diagonal elements are all equal to 1, then the Tucker3 model may be simplified to a PARAFAC model.

The loadings matrices are estimated through solving algorithms. Some examples of these algorithms are Alternating least squares (ALS) [29], Alternating Slice-Wise Diagonalization (ASD) [30], Positive Matrix Factorization for 3-way arrays (PMF3)

[31], and finally direct and non-iterative algorithm as Generalized Rank Annihilation Method (GRAM) [32].

5.3 Paper: An extension of PARAFAC to analyze multi-group three-way data

Rotari Marta, Valeria Fonseca Diaz, Bart De Ketelaere and Murat Kulahci, *"An extension of PARAFAC to analyze multi-group three-way data"*, under revision at the Chemometrics and Intelligent Laboratory Systems Journal.

An extension of PARAFAC to analyze multi-group three-way data

Marta Rotari[a], Valeria Fonseca Diaz[c], Bart De Ketelaere[c], Murat Kulahci[a,b]

[a]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark*
[b]*Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden*
[c]*Department of Biosystems, MeBioS division, KU Leuven, Leuven, Belgium*

Abstract

This paper introduces a novel methodology for analyzing three-way array data with a multi-group structure. Three-way arrays are commonly observed in various domains, including image analysis, chemometrics, and real-world applications. In this paper, we use a practical case study of process modelling in additive manufacturing, where batches are structured according to multiple groups. Vast volumes of data for multiple variables and process stages are recorded by sensors installed on the production line for each batch. For these three-way arrays, the link between the final product and the observations creates a grouping structure in the observations. This grouping may hamper gaining insight into the process if only some of the groups dominate the controlled variability of the products. In this study, we develop an extension of the PARAFAC model that takes into account the grouping structure of three-way data sets. With this extension, it is possible to estimate a model that is representative of all the groups simultaneously by finding their common structure. The proposed model has been applied to three simulation data sets and a real manufacturing case study. The capability to find the common structure of the groups is compared to PARAFAC and the insights into the importance of variables delivered by the two models are discussed.

Keywords: PARAFAC; Factor analysis; Additive manufacturing

Introduction

Manufacturing organizations strive to adopt innovative production methods to enhance their efficiency and design flexibility. However, modern production techniques can be complex and not yet fully understood. Achieving a complete comprehension and optimization of these processes requires identifying the variables and components of the process, as well as the conditions under which production levels can be optimized. To achieve such an understanding, it is essential to develop descriptive and interpretable data-driven models that provide clear insights into the data.

Modern manufacturing industrial systems are often equipped with technology that can collect data from a large number of sources, resulting in a voluminous amount of information. This can produce datasets that are structured in three-way arrays, organized in observations collected for multiple variables and several conditions. For example, in a batch fermentation process, data collected in time for multiple variables and multiple batches will be in the form of a three-dimensional array [1, 2]. A second example is processes based on sensory data, which may come from different channels and multiple devices simultaneously [3, 4, 5].

Several approaches have been proposed to model three-way data using supervised or unsupervised methods, ranging from linear models [6, 7, 8, 9, 10] to black-box models for tensors [11, 12]. However, to find the optimal settings of the system, it is required to fully understand the conditions and the interactions among the variables involved in the production processes. Linear models remain a predominant tool for three-way data as they can identify key factors that impact the process and guide decision-making. Although black-box models may offer high prediction accuracy, they lack the interpretability and transparency necessary to understand the system.

The most predominant methodologies within the class of linear models for unsupervised three-way data have been factor analysis methods such as PARAFAC [8, 13] and Tucker3 model [9, 14]. These models are used by analyzing the covariance structure among the variables and among different experimental conditions providing the so-called loadings matrices whose values are indicative of the importance of each variable and each condition in the total variability of the three-way data [15, 2]. The scores that are delivered for the observations of the data allow for the

identification of the dispersion among the samples and the differences in the variance explained by each component of the model.

In many industrial processes, the existence of sub-groups within the observations is a common occurrence that can greatly impact the analysis and interpretation of the results. The multi-group structure in the data refers to the presence of distinct groups that may exhibit unique characteristics and behaviours. Groups of observations are likely to happen when the observations or samples share a specific characteristic that identifies the group [16, 17, 18]. Various factors, such as demographic variables, geographic location, or other relevant criteria, may define these groups. In the manufacturing processes, the observations collected during production are linked to the final products, which individuate a multi-group structure of the observations. Understanding the multi-group structure in the data is crucial for accurate analysis, as failing to consider these groups can lead to biased results and erroneous conclusions. Thus, it is important to carefully identify and account for the multi-group structure in the model in order to ensure that research findings are valid and applicable to all relevant groups [19, 18].

When grouping structures exist, the sources of variability explained in the resulting model can be dominated by specific groups without representing other groups' variability. Traditional models for three-way arrays that do not take into account the group structure can be affected by the greater variance of a dominant group and fail to represent the entire data set. Therefore, the importance of specific variables or conditions that are extracted from the model parameters may not hold for all the groups simultaneously, leading to potentially erroneous conclusions. One practical solution to take into account the presence of multiple groups is to analyze the groups separately. However, this strategy would result in an abundance of parameters and the lack of a single unifying model of the complete system.

Linear models for high-dimensional data considering the grouping structure have been widely studied for the case of two-way data [20, 21, 18, 22, 23]. It has been shown how the resulting model, when considering the group structure of the observations, can be informative in understanding the common cause of variability across all groups [16, 24, 25]. Through multi-group models, it is possible to interpret the functionality of the production systems for several groups simultaneously.

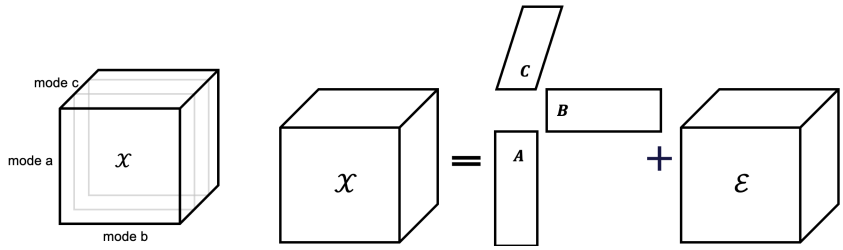
The current study presents the development of an extension of the PARAFAC method that adjusts the model parameters according to the multi-group structure of the data. The model aims to identify a common structure by calculating common loadings in the presence of multi-groups in the three-way data. By identifying common loadings, the proposed model provides a more accurate and comprehensive description of the underlying process. The article is organized as follows. First, the extended PARAFAC method is presented with the proposed algorithm to fit the model. Then, the case studies with simulated and real data are presented, highlighting the multi-group structure for three-way data. The results of the PARAFAC and the extended PARAFAC models are presented and compared. The models are used to analyze the most important variables and conditions in the datasets delivered by each model. Finally, the conclusions are presented.

1. Methods

This section provides a brief description of the PARAFAC model and an algorithm for its solution, Alternating Least Squares (ALS). Following that, we present the proposed extension aimed to handle a multi-group structure. Three-way arrays are denoted by $\mathcal{X} \in \mathbb{R}^{N\times M\times K}$ whose elements are x_{ijk} where $i = 1, \cdots, N$, $j = 1, \cdots, M$ and $k = 1, \cdots, K$. The three-way entries are denoted by modes a, b and c as shown in Figure 1a. In this structure, mode a represents the observations, mode b represents the variables, and mode c represents different conditions, such as time, different levels of temperatures, etc. The elements that belong to modes a and b with mode c fixed will be denoted by $\mathcal{X}_{\bullet\bullet k} = (x_{ijk})$ where k is fixed, $i = 1, \cdots, N$ and $j = 1, \cdots, M$. To define matrices, [,] is used for horizontal concatenation and [;] for vertical concatenation. A matrix or two-way array is denoted by a capital bold letter $\mathbf{X} = [\mathbf{X}_{\bullet 1}, \ldots, \mathbf{X}_{\bullet j}, \ldots, \mathbf{X}_{\bullet M}] = [\mathbf{X}_{1\bullet}^T; \ldots; \mathbf{X}_{i\bullet}^T; \ldots; \mathbf{X}_{N\bullet}^T]$ with $\mathbf{X}_{\bullet j}$ representing its j–the column and $\mathbf{X}_{i\bullet}^T$ its i-th row. The letters N, M, K are reserved for indicating the dimensions and F for indicating the number of components. Superindex ^{T} represents the transpose operator.

1.1. The PARAFAC method

The PARAFAC method is a decomposition method of the three-way array into three matrices: the score matrix $\mathbf{A}$ and two loadings matrices $\mathbf{B}$ and $\mathbf{C}$. Matrix $\mathbf{A}$ is referred to as scores as its entries represent the numerical value of the



(a) Three-way array $\mathcal{X}$.

(b) PARAFAC model representation.

Figure 1: (*a*): Three-way array $\mathcal{X}$ represented by the modes *a*, *b* and *c*. (*b*): PARAFAC model representation, $\mathcal{X}$ three-way array, **A**,**B** and **C** model loadingss matrices and $\mathcal{E}$ three-way residual array.

observations [8, 15]. Matrices **B** and **C** are commonly referred to as loadings as they represent the numerical value of variables and conditions, respectively. Considering a three-way array $\mathcal{X}$, a PARAFAC model with F components can be written as follows:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{1}$$

where $\mathbf{A} = (a_{if})$ is the score matrix $\mathbb{R}^{N\times F}$, $\mathbf{B} = (b_{jf}) \in \mathbb{R}^{M\times F}$ is the loadings matrix of mode *b*, and $\mathbf{C} = (c_{kf}) \in \mathbb{R}^{K\times F}$ is the loadings matrix of mode *c*. A visual representation of the PARAFAC model is shown in Figure 1b. In particular, the reconstructed three-way array is obtained as the outer product $\otimes$ of the three matrices $\sum_{f=1}^{F} \mathbf{A}^f \otimes \mathbf{B}^f \otimes \mathbf{C}^f$ plus the residuals array $\mathcal{E} = (e_{ijk}) \in \mathbb{R}^{N\times M\times K}$. The PARAFAC model results from finding the model matrices that minimize the sum of squares of the residuals:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{ijk} (x_{ijk} - \sum_{f=1}^{F} a_{if} b_{jf} c_{kf})^2 = \min \sum_{ijk} (e_{ijk})^2 \tag{2}$$

Several algorithms have been proposed to solve Eq. (2) for **A**, **B**, and **C** such as Alternating Least Squares (ALS) [8, 26, 27], Derivative Computations [28] and direct (non-iterative) procedures [29, 30]. We present here the ALS algorithm as it is the original and most widely used method to estimate the PARAFAC model [8].

ALS algorithm

The ALS algorithm aims at calculating the model matrices **A**, **B** and **C** that minimize the sum of squared residuals in Eq. (2) using an iterative process. At each step *t*, it calculates and updates the matrices until convergence or until the maximum number of iterations is reached rendering a series of matrices $\{\mathbf{A}^{(t)}\}$, $\{\mathbf{B}^{(t)}\}$ and $\{\mathbf{C}^{(t)}\}$ [8, 26, 27]. The first step consists in initializing the matrices $\mathbf{B}^{(1)}$ and $\mathbf{C}^{(1)}$ as random matrices. To calculate the score matrix $\mathbf{A}^{(t)}$, it is required to first unfold the three-way array $\mathcal{X}$ in a two-way array as $\mathbf{X_a} = [\mathcal{X}_{\bullet\bullet 1}\ ,\ \ldots\ ,\ \mathcal{X}_{\bullet\bullet K}]$ of dimensions $N \times MK$. The matrix $\mathbf{A}^{(t)}$ is calculated by solving the minimization problem

$$\min_{\mathbf{A}} \|\mathbf{X_a} - \mathbf{A}^{(t)}\mathbf{Z}^T\| \ \text{ where } \ \mathbf{Z} = \mathbf{B}^{(t-1)} \otimes \mathbf{C}^{(t-1)} \tag{3}$$

Matrices $\mathbf{B}^{(t)}$ and $\mathbf{C}^{(t)}$ are calculated in an analogous way. To calculate $\mathbf{B}^{(t)}$, $\mathcal{X}$ is unfolded as the matrix $\mathbf{X_b} = [\mathcal{X}_{1\bullet\bullet}\ ,\ \ldots\ ,\ \mathcal{X}_{N\bullet\bullet}]$ of dimensions $M \times KN$. For $\mathbf{C}^{(t)}$, $\mathcal{X}$ is unfolded as the matrix $\mathbf{X_c} = [\mathcal{X}_{\bullet 1\bullet}\ ,\ \ldots\ ,\ \mathcal{X}_{\bullet M\bullet}]$ of

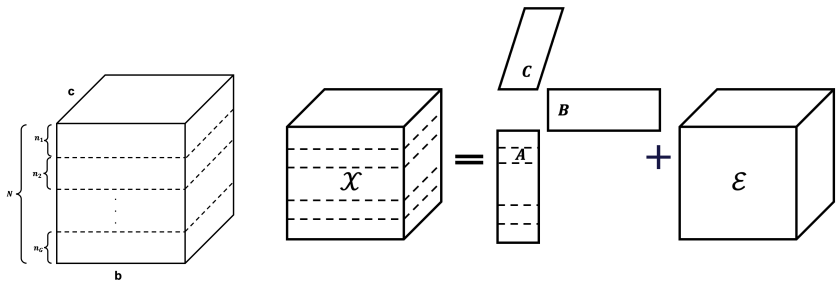
dimensions $K \times NM$. The new updated matrices $\mathbf{B}^{(t)}$ and $\mathbf{C}^{(t)}$ solve the following problems:

$$\min_{\mathbf{B}} \|\mathbf{X_b} - \mathbf{B}^{(t)}\mathbf{Z}^T\| \text{ where } \mathbf{Z} = \mathbf{A}^{(t)} \otimes \mathbf{C}^{(t-1)},\ \mathbf{X_b} \in \mathbb{R}^{M \times KN}$$

and

$$\min_{\mathbf{C}} \|\mathbf{X_c} - \mathbf{C}^{(t)}\mathbf{Z}^T\| \text{ where } \mathbf{Z} = \mathbf{A}^{(t)} \otimes \mathbf{B}^{(t)},\ \mathbf{X_c} \in \mathbb{R}^{K \times NM}$$

1.2. Proposed method for multi-group data



(a) Three-way array $\mathcal{X}$ with multi-group structure.

(b) Proposed model representation.

Figure 2: (*a*): Multi-group representation of the three-way array $\mathcal{X} = x_{ijk} \in \mathbb{R}^{N \times M \times K}$ which present a multi-group structure. Along the mode a the observations are divided into G groups, such that group $g = \{1, \cdots, G\}$ is represented by $1, \cdots, n_g$ rows and $\sum_{g=1}^{G} n_g = N$. (*b*): Representation of the new proposed model.

In a multi-group structure, the three-way array $\mathcal{X}$ is composed by the same three modes a, b, and c, when the observations in mode a are are divided into G groups. A visual representation of this structure is shown in Figure 2a. Each group $g \in \{1, \cdots, G\}$ is represented by $1, \cdots, n_g$ rows, such that $\sum_{g=1}^{G} n_g = N$. The sub-array that contains the observations belonging to group g is represented by $\mathcal{X}_g \in \mathbb{R}^{n_g \times M \times K}$ where n_g is the number of observations in such group. The two main characteristics of the proposed method are:

- **A**, **B**, and **C** are estimated from the decomposition of the three-way array $\mathcal{X}$, constrained by its group structure.
- The resulting loadings matrices **B** and **C** focus on the common variability among the groups.

In this way, the resulting model is not dominated by a specific group but is representative of all the groups uniformly. The optimization function of the proposed method is defined as

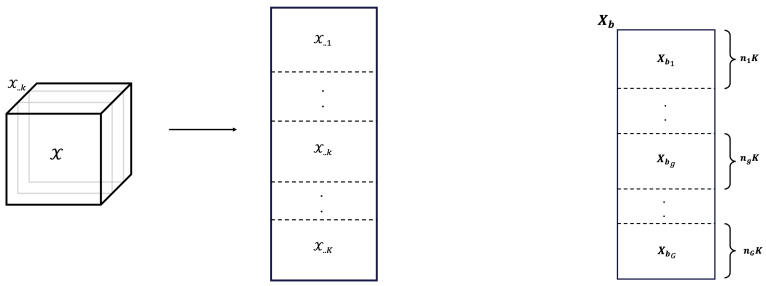
$$\min_{\mathbf{A}_g, \mathbf{B}, \mathbf{C}} \sum_{ijk} (x_{ijk} - \sum_{f=1}^{F} a_{g_{if}} b_{jf} c_{kf})^2$$
$$\text{such that } \sum_{g=1}^{G} \|\mathbf{B}_g - \mathbf{B}\|^2 < \gamma_b \text{ and } \sum_{g=1}^{G} \|\mathbf{C}_g - \mathbf{C}\|^2 < \gamma_c \tag{4}$$

where $\mathbf{A}_g = (a_{i_g f})$ is the score matrix containing the observations of matrix **A** that belong to the group g where $i_g = 1, \cdots, n_g$, $\mathbf{B}_g$ and $\mathbf{C}_g$ are the loadings matrices corresponding to the observations in group g, and **B** and **C** are the common loadings matrices across all groups in $\mathcal{X}$. We present here the extension of the ALS algorithm to solve the problem presented in Eq. (4).

Extended ALS algorithm

Similar to the ALS algorithm for the PARAFAC model, the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are updated at each step of the algorithm by minimizing the sum of squares of residuals between the input matrix $\mathcal{X}$ and the outer product of the three model matrices. Thus, we shall generate a series of matrices $\{\mathbf{A}^{(t)}\}$, $\{\mathbf{B}^{(t)}\}$ and $\{\mathbf{C}^{(t)}\}$ that reach convergence at each successive step. However, the calculation of $\mathbf{B}$ and $\mathbf{C}$ will be influenced by the loading matrices of each group.

The algorithm is initialized by calculating the PARAFAC model with solution denoted by $\mathbf{A}^{(1)}, \mathbf{B}^{(1)}, \mathbf{C}^{(1)}$. Therefore, in the absence of groups, this extension returns to the conventional PARAFAC model. In the presence of multiple groups, the algorithm first solves the PARAFAC model for each group and then finds the common solution across all groups. We describe here the process of calculating $\mathbf{B}$ at step t. For this, the three-way array $\mathcal{X}$ is converted into an unfolded two-way array $\mathbf{X_b}$, defined as $[\mathcal{X}_{\bullet\bullet 1} \ ; \ \dots \ ; \ \mathcal{X}_{\bullet\bullet K}] \in \mathbb{R}^{NK\times M}$ (See Figure 3a).



(a) The unfolding of the three-way array $\mathcal{X}$.

(b) Representation of the matrix $\mathbf{X_b}$.

Figure 3: (a): The unfolding of the three-way array $\mathcal{X}$ into the two-way array $[\mathcal{X}_{\bullet\bullet 1} \ ; \ \dots \ ; \ \mathcal{X}_{\bullet\bullet K}]$. (b): Representation of the rearranged unfolded matrix $\mathbf{X_b}$ with the multi-group structure $\mathbf{X_b} = [\mathbf{X_b}_1 \ ; \ \dots \ ; \ \mathbf{X_b}_G]$.

We reorganize the rows of $\mathbf{X_b}$ by grouping the observations that belong to the same group. This generates $\mathbf{X_b}$ with a multi-group structure as $\mathbf{X_b} = [\mathbf{X_b}_1 \ ; \ \dots \ ; \ \mathbf{X_b}_G]$ where each group is of dimensions $n_gK \times M$, as represented in Figure 3b. For each group thus created, we calculate the group loadings matrix $\mathbf{B}_g^{(t)}$ by solving the objective function:

$$\min_{\mathbf{B}_g} \|\mathbf{X_b}_g - \mathbf{Z}_g\mathbf{B}_g^{(t)T}\| \ \text{ where } \ \mathbf{Z}_g = \mathbf{C}^{(t-1)} \otimes \mathbf{A}_g^{(t-1)} \tag{5}$$

where $\mathbf{A}_g^{(t-1)}$ denotes the rows of the matrix $\mathbf{A}^{(t-1)}$ that correspond to the group g. The procedure is repeated for $g = 1, \cdots, G$ obtaining the group loadings matrices $\{\mathbf{B}_1^{(t)}, \cdots, \mathbf{B}_G^{(t)}\}$. The matrix $\mathbf{B}^{(t)}$ corresponds to the common loadings matrix as represented in Figure 4. These common loadings $\mathbf{B}^{(t)}$ are computed using a weighted mean. Because the model needs to represent all groups uniformly, these weights are determined based on the variance explained for each group. If a group has a high variance explained, its weight will be lower. Correspondingly, the groups with lower variance explained will have a higher weight. This renders a common loadings matrix represented by variance homogeneity [22, 23].

The calculation of $\mathbf{C}_g$ and $\mathbf{C}$ at step t follows the same process as before. The three-way array $\mathcal{X}$ is unfolded as $\mathbf{X_c} = [\mathcal{X}_{\bullet 1\bullet} \ ; \ \dots \ ; \ \mathcal{X}_{\bullet M\bullet}] \in \mathbb{R}^{NM\times K}$. The observations are reorganized so that $\mathbf{X_c}$ shows a multi-group structure matrix. For each group $g = 1, \cdots, G$ we calculate the corresponding group loadings matrix $\mathbf{C}_g^{(t)}$ by solving the problem given by:

$$\min_{\mathbf{C}_g} \|\mathbf{X_c}_g - \mathbf{Z}_g\mathbf{C}_g^{(t)T}\| \ \text{ where } \ \mathbf{Z}_g = \mathbf{A}_g^{(t-1)} \otimes \mathbf{B}^{(t)} \tag{6}$$

Here, the algorithm uses the matrix $\mathbf{A}^{(t-1)}$ of the previous iteration and the latest updated loadings $\mathbf{B}^{(t)}$. For each group $g = 1, \cdots, G$ we calculate the group loadings matrices $\{\mathbf{C}_1, \cdots, \mathbf{C}_G\}$ and the matrix $\mathbf{C}^{(t)}$ becomes the common loadings matrix to all the groups obtained by weighted mean.

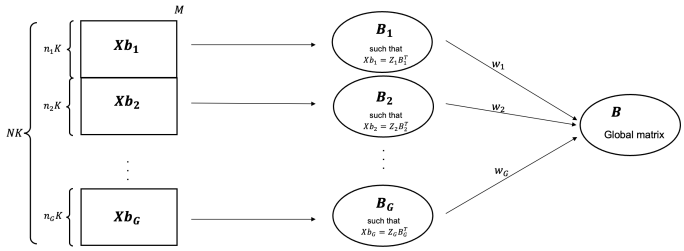


Figure 4: Representation of the calculation of a matrix common to all groups in a generic step i of the algorithm. This scheme applies also to the calculation of $\mathbf{C}$.

Finally, to update the score matrix $\mathbf{A}$ we unfold the three-way array $\mathbf{X}$ into the $N \times MK$ matrix $\mathbf{X}_a$ as in the case of PARAFAC and solve the same PARAFAC step for each group g to obtain $\{\mathbf{A}_1^{(t)}, \ldots, \mathbf{A}_G^{(t)}\}$. Finally we repeat the above steps until convergence or the maximum number of iterations is reached. Algorithm 1 describes the entire procedure of the proposed method.

The weighting scheme was determined based on the optimization problem of Eq. (4). The solution to the constraints to determine a common $\mathbf{B}$ and $\mathbf{C}$ can be obtained in several ways. One option is to minimize each constraint, in which case the solution is given by a simple mean over the groups. A second option would consist in including the constraint within each unfold optimization. In this case, the algorithm would need regularization parameters (i.e. Lagrange multipliers) which would add extra tuning parameters to estimate the model. For the sake of limiting the complexity of building the model while complying with the objective of finding the common loadings across groups, a weighted average was embedded in the algorithm where the weights correspond to the inverse of the variance of each group explained within the model.

Algorithm 1 Extended ALS algorithm for PARAFAC model with multi-group data

Input: $\mathcal{X}$ data, list G group identification, F number of components
  $\mathbf{A}, \mathbf{B}, \mathbf{C} \leftarrow$ parafac$(\mathcal{X}, F)$
  group-weights $= 1 - var\_explained(1, \cdots, G)$
  group-weights $= \frac{group-weights}{\|group-weights\|}$
  while convergence or max_iterations **do**
    for $g = 1, \cdots, G$ **do**
      $\mathbf{Z}_g = \mathbf{C} \otimes \mathbf{A}_g$
      $\mathbf{B}_g = \min_B \|\mathbf{X}_{\mathbf{b}g} - \mathbf{Z}_g \mathbf{B}_g^T\|$
    end for
    For each model component $f = 1, \cdots, F$
    $\mathbf{B} = weighted\_mean(\mathbf{B}_1, \cdots, \mathbf{B}_G)$
    for $g = 1, \cdots, G$ **do**
      $\mathbf{Z}_g = \mathbf{B} \otimes \mathbf{A}_g$
      $\mathbf{C}_g = \min_C \|\mathbf{X}_{\mathbf{c}g} - \mathbf{Z}_g \mathbf{C}_g^T\|$
    end for
    For each model component $f = 1, \cdots, F$
    $\mathbf{C} = weighted\_mean(\mathbf{C}_1, \cdots, \mathbf{C}_G)$
    for $i = 1, \cdots, G$ **do**
      $\mathbf{Z} = \mathbf{B} \otimes \mathbf{C}$
      $\mathbf{A}_i = \min_{A_g} \|\mathbf{X}_{\mathbf{a}g} - \mathbf{A}_g \mathbf{Z}^T\|$
    end for
    $\mathbf{A} = [\mathbf{A}_1, \cdots, \mathbf{A}_G]^T$
  end while
Output: $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$

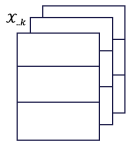
2. Data Overview

To evaluate the performance of the proposed model, three simulated data sets and one real case study from additive manufacturing were used. Each simulation reflects a different scenario for the grouping structure of three-way arrays. The first simulation involves simulating data in batches for $k = 1, 2, 3$, where each batch is composed of three groups. The second and third simulations use model loadings $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ to generate the three-way array. Various shifts in group variance, observations and variables were introduced to test the algorithm's performance. In the real case study, we examined a three-way array $\mathcal{X}$ that represents three batches from an additive manufacturing process.

The PARAFAC model and the proposed extension were fitted to each dataset. The two methods were compared based on their ability to homogenise the variance across groups, create common loadings for all the groups and provide informative insights through variable loadings. Our goal was to demonstrate the efficacy of the proposed approach in accurately analysing complex multi-group array data and identifying a common model. All the analyses were performed in R 4.0 with in-house codes.

2.1. Data: Simulated data 1

The first simulation defines a three-way array $\mathcal{X} \in \mathbb{R}^{(90\times40\times3)}$ consisting of three groups of observations. In this simulation we define the matrices $\mathcal{X}_{\bullet\bullet k}$ $(k = 1, 2, 3)$, each composed of three groups. Each group of each matrix is simulated using a multivariate normal distribution $N_{40}(\cdot, \cdot)$ with mean $\mu = 0$ and standard deviations $\sigma = 2, 5, 1$ as shown below:



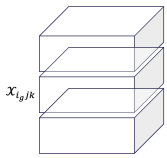
$$\mathcal{X}_{\bullet\bullet k_{g=1}} \sim N_{40}(0, 2\mathbf{I})$$
$$\mathcal{X}_{\bullet\bullet k_{g=2}} \sim N_{40}(0, 5\mathbf{I})$$
$$\mathcal{X}_{\bullet\bullet k_{g=3}} \sim N_{40}(0, \mathbf{I})$$

where $\mathbf{I}$ represents the identity matrix of the corresponding order. After concatenating the groups and the matrices $\mathcal{X}_{\bullet\bullet k}$, an error with a normal distribution of $\mu = 0$ and standard deviation $\sigma = 0.05$ was added to obtain the final array $\mathcal{X}$.

It is worth noticing that the second group, $\mathcal{X}_{\bullet\bullet k_{g=2}}$, has been assigned a larger standard deviation to challenge the ability of the proposed model to homogenise the variances across the groups. This will test the algorithm's capability to overcome any potential biases towards a particular group with larger variation and to generate informative loadings that reflect the entire three-way array.

2.2. Data: Simulated data 2

The second simulation consists of a three-way array $\mathcal{X} \in \mathbb{R}^{(90\times20\times3)}$ with a multi-group structure consisting of three groups as well. In this case, we simulated each sub-array $\mathcal{X}_g \in \mathbb{R}^{n_g\times20\times3}$ for $n_g = 30, 30, 30$ and concatenate them vertically to obtain the final three-way array. Each sub-array $\mathcal{X}_g$ is obtained as the outer product of matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ generated using the Normal distribution for $\mathbf{A}$ and Uniform distribution for $\mathbf{B}$ and $\mathbf{C}$.



$$\mathbf{A} \sim N(0, \mathbf{I})$$
$$\mathbf{B} \sim U(0, \mathbf{I})$$
$$\mathbf{C} \sim U(0, \mathbf{I})$$

Within the established framework, a shift in the model matrix was introduced to create separation between groups. The aim is to examine the effect of the shifts on both models. In this case, a shift was introduced in the score matrix $\mathbf{A}$ corresponding to the first and third groups. Specifically, $\mathbf{A}_{\bullet 3_{g=1}}$ was multiplied by factor of 4 and $\mathbf{A}_{\bullet 3_{g=3}}$ was multiplied by -1. After the concatenation of the groups, an error with a normal distribution of $\mu = 0$ and standard deviation $\sigma = 0.05$ was added to obtain the final array.

2.3. Data: Simulated data 3

A third simulation is obtained using a similar framework as in simulated data 2 of Section 2.2. We generate a three-way array $\mathcal{X} \in \mathbb{R}^{(90\times20\times3)}$, vertical concatenation of three sub-arrays $\mathcal{X}_g \in \mathbb{R}^{n_g\times20\times3}$ for $n_g = 30, 30, 30$. Each sub-way array $\mathcal{X}_g$ is obtained as the outer product of matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, as in the previous example.

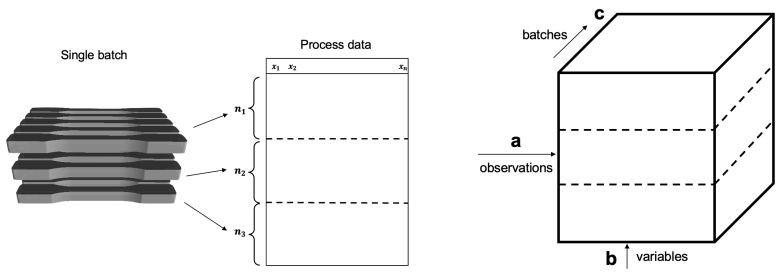
Here, a shift in the loadings matrix $\mathbf{B}$ was introduced. Specifically, the first column of the first group was multiplied by 5, (i.e. $5\mathbf{B}_{\bullet 1_{g=1}}$). After the concatenation of the groups, an error with a normal distribution of $\mu = 0$ and standard deviation $\sigma = 0.05$ was added to obtain the final array.

The objective of this simulation study is to test the capability of the proposed model in generating common loadings that mitigate the variations unique to each group, thereby enabling the recovery of common information shared across all groups. We aim to showcase that the model captures the underlying common patterns while minimizing the impact of group-specific variations.

2.4. Case study: Additive manufacturing

The real case study corresponds to an additive manufacturing process for high-volume 3D printing. The so-called Selective Thermoplastic Electrophotographic Process (STEP) [31] takes place by rendering a three-dimensional object

from a digital model. This process generates a 3D bulk structure by fusing and pressing super-thin layers. Multiple sensors are positioned throughout the production chain on the new manufacturing line measuring several variables for each super-thin layer.



(a) Representations of the multi-group structure.

(b) Three-way array representation in the Additive manufacturing studying case.

Figure 5: (*a*): The creation of a multi-group structure in the process data as a result of the relationship between the batch products and the observations in the process data. (*b*) Observations, variables and batches organisation in the three-way array in the Additive manufacturing case study.

Our study focuses on the analysis of a data set consisting of four batches that form a three-way array denoted as $\mathcal{X}$. In the $\mathcal{X}$ three-way array, mode b contains 53 continuous variables collected through the sensors located on the printing machine. The variables are labelled from $V1$ to $V53$ due to confidentiality. Mode a represents the observations of the variables collected for each super-thin layer. Finally, mode c represents the different batches, as shown in Figure 5b. Each batch consists of three final groups of products. The association between the final products and the observations creates a multi-group structure of three groups represented in the $\mathcal{X}$ three-way array, Figure 5a. The three-way array was previously scaled by group [32]. Group scaling involves the scaling of each variable within the sub-array $\mathcal{X}_g$ based on the mean and standard deviation calculated within the respective group denoted as $\mathcal{X}_{\bullet j \bullet_g}$ for $g \in 1, \cdots, G$. This approach ensures that variables within each group are standardized in relation to the group-specific distribution characteristics.

This case study aims at gaining insights into the covariance structure among the variables and the different batches. Specifically, we aim to analyze the loadings matrix $\mathbf{B}$, which reveals the relative importance of each variable in explaining the total variability of the three-way array. To accomplish this, we employ both the PARAFAC model and the proposed extension.

3. Results and Discussion

As a rule of thumb, we present all the results across all data sets using models with only 10 components. We adopted this choice to effectively showcase two key objectives. Firstly, through Simulated data set 1, our aim was to demonstrate the consistency of the explained variance across all groups. Additionally, Simulated data sets 2 and 3 were conducted to provide further evidence of the proposed model's ability to generate common loadings that are shared among all groups. It is important to note, however, that the determination of the number of components is context-dependent and should be tailored to the specific characteristics of each individual case, and may vary accordingly.

Simulated data sets

The percentage of variance explained by both models for each group in Simulation 1 is summarized in Table 1. The PARAFAC model reveals that the second group has a higher representation than the other two groups in all of the components. This result is in line with the ranking of the standard deviations set for each group as presented in Section 2.1.

Table 1: Cumulative % of Variance Explained by Groups in Simulated dataset 1.

PARAFAC Algorithm										
	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
Group1	1.26	2.76	3.71	5.10	6.25	7.89	8.60	10.74	12.40	14.15
Group2	3.61	7.43	11.33	14.72	18.44	21.37	25.28	27.71	31.26	33.71
Group3	1.15	2.33	3.44	4.19	4.91	5.24	6.61	7.76	8.68	10.18
Proposed Algorithm										
	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
Group 1	2.27	4.45	6.39	9.05	9.42	12.93	15.71	16.65	17.98	20.94
Group2	2.40	4.79	7.00	8.50	10.41	13.17	15.67	18.00	18.67	21.60
Group3	2.63	4.82	6.71	9.50	11.30	13.82	14.90	16.83	17.78	22.04

To compare the representation of the groups across all of the components, the ratio of the explained variances for each group with respect to group 3 was calculated (See Figure 6). Throughout the different components, the second group dominated the representativity with a percentage of explained variance 3 times the variability group 3. This result showed the dominance of the second group for the resulting PARAFAC model with a lower representation of the other groups. This is indicative of a model that represents only one group rather than being representative of the entirety of the data. In contrast, the proposed model demonstrates a uniform explained variance across all 3 groups, providing evidence of a model that uniformly represents the entire dataset. This is supported by the ratios in Figure 6, which depict a more uniform explained variance across all groups throughout the components.

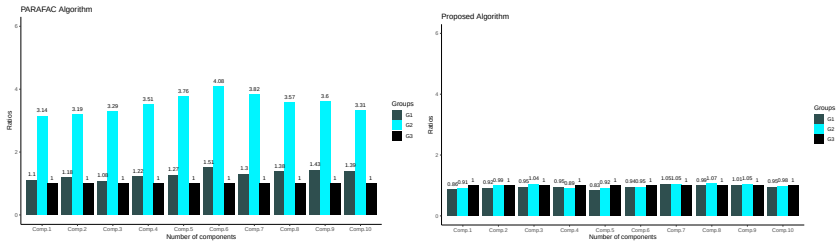


Figure 6: The ratios of the explained variances for each group with respect to group 3 in simulated data set 1.

Figure 7 displays the scores of the 1st and 2nd components for Simulation 2, described in Section 2.2. In this case, a shift in the score matrix **A** was applied in the data simulation. The analysis of the results of the two models, therefore, focuses on the visualization of the first two scores.

In the case of the PARAFAC model, while the scores of the second and third groups nearly overlap and vary along the same direction, the scores of the third group are almost perpendicular to those of the first two groups, with a much larger dispersion. In contrast, the proposed model demonstrates nearly overlapping scores across all groups. Additionally, it can be noticed that the direction of variation for all groups is closer to one another. This provides evidence that the proposed model effectively mitigates the inherent shifts between groups and captures shared directions.

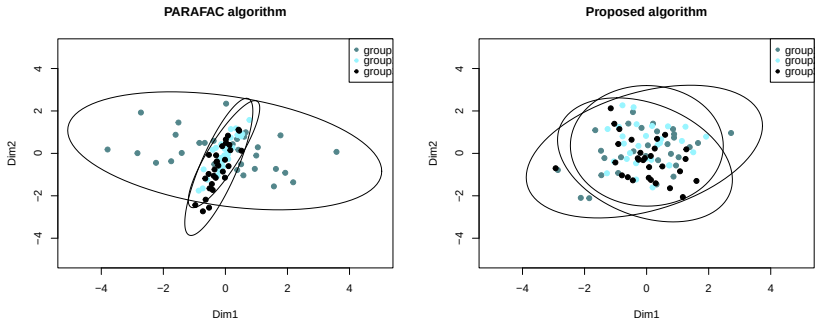


Figure 7: First two scores plot for the PARAFAC model and the Proposed model with a 99% confidence interval ellipse related to Simulation 2, described in Section 2.2.

In Figure 8, the **B** loadings matrices are presented for both the PARAFAC and proposed models referred to Simulation 3, described in Section 2.3. The black profile in the figures represents the original loadings matrix, and a significant shift is highlighted represented by the first group. This extreme shift scenario was deliberately chosen to assess the algorithm's performance under challenging conditions. However, in practical scenarios, the variations between groups are expected to be comparatively smaller.

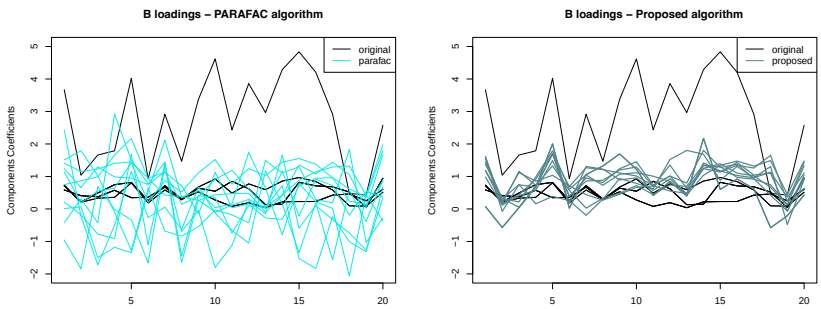


Figure 8: The PARAFAC and proposed model loadings matrices **B** for the dataset 3, Section 2.3.

Upon comparing the results of the two models, noticeable differences can be observed. The PARAFAC model exhibits a greater bias towards the shift, resulting in **B** loadings profiles with higher levels of noise. These loadings exhibit extremely high peaks, influenced by the larger variability present in the first group. Conversely, the proposed model displays less pronounced peaks and generally smoother trend profiles. The loadings of the proposed model are closer to those of groups 2 and 3, indicating reduced bias from group 1.

These findings suggest that the proposed model can identify the common patterns present in all groups while mitigating any bias towards a specific group. Overall, these results highlight the potential of the proposed model to

serve as a reliable tool for analyzing multi-group datasets.

Case study: Additive manufacturing

The results of the PARAFAC and the proposed models in terms of the explained variance per group for the three-way array $\mathcal{X}$ are summarized in Table 2. Consistent with the findings observed in the simulated datasets, the PARAFAC model exhibited a relatively higher representation of the first group compared to the subsequent two groups. This disparity in groups representation becomes more pronounced, particularly in the higher-order components. In contrast, the proposed model, resulted in a more uniform representation of all three groups.

Table 2: Cumulative % of Variance Explained by Groups: Low-quality products $\mathcal{X}$ three-way matrix

PARAFAC Algorithm										
	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
Group 1	39.00	39.25	39.18	53.57	53.76	65.48	66.00	66.22	67.93	71.53
Group2	26.85	26.90	27.49	33.15	35.74	39.58	42.55	43.15	48.79	52.57
Group3	20.66	21.42	21.70	27.17	44.48	46.98	47.69	48.47	52.79	55.54
Proposed Algorithm										
	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9	Comp 10
Group 1	32.90	33.10	33.30	38.27	42.62	46.53	50.30	56.59	59.70	58.36
Group2	29.71	29.69	30.18	36.88	41.08	39.33	47.20	48.31	53.68	53.06
Group3	23.38	24.21	24.65	30.06	36.84	36.30	44.05	48.76	52.84	55.59

These findings are further validated by the ratios between the explained variances of each group with respect to group 3, in Figure 9. The initial components show that the PARAFAC model exhibits a two-fold ratio in representing the first group when compared to the third group. While the disparity diminishes with subsequent components, a preference for the first group persists, albeit to a lesser degree. This observation suggests that the PARAFAC model captures the variability of the first group more effectively than it does for the entire dataset. In contrast, the proposed model already explains similar degree of variances explained by all groups in the first few components. As more components are added, the model portrays an equal representation of all three groups. This demonstrates that the proposed model is more effective at capturing the common variability and fostering a broader understanding of the data.

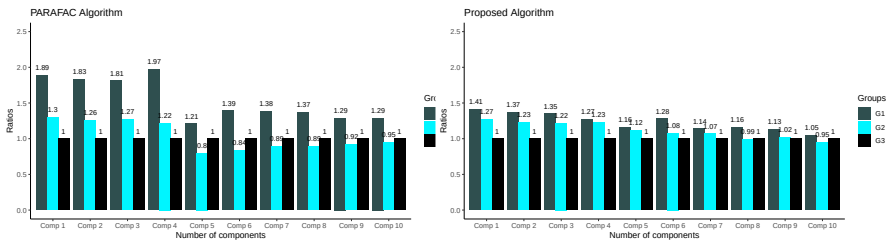


Figure 9: Case study: Ratios of the explained variance for each group with respect to group 3.

Figure 10 displays the scores of the first two components for both models. In the PARAFAC case, it is evident that the scores corresponding to group 1 exhibit greater separation than those of the remaining groups, in agreement

with the above findings. Moreover, the direction of the scores of group 2 is different from the other two groups. Conversely, in the proposed model scenario, the scores of all three groups exhibit significant overlap within a common region. Notably, the groups also demonstrate shared patterns direction of variability. This outcome is consistent with the objective of the proposed model, which aims to recover loadings that express the shared patterns and variability across all groups.

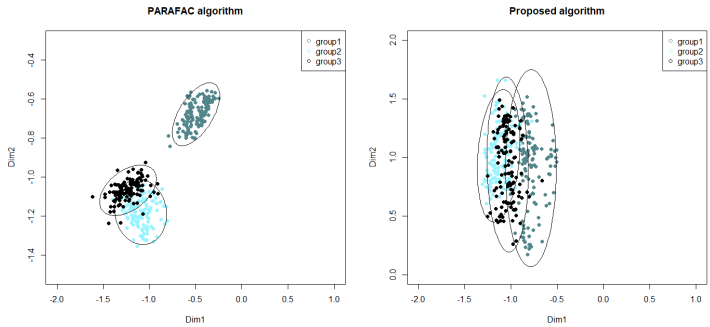


Figure 10: Case study: Low-quality products three-way array $\mathcal{Y}$. First two dimensions scores representation with a 99% confidence interval ellipse.

The biplot of the loadings matrices **B** for both models are illustrated in Figure 11. In the PARAFAC model, the variables $V2$, $V4$, $V8$, $V11$, $V44$ and $V45$ have the highest model coefficients. The proposed model showcases similar high coefficients for the variables $V11$, $V44$ and $V45$, albeit with slightly reduced magnitudes. Moreover, other variables such as $V1$, $V2$, $V8$, $V9$, and $V10$ are shown to have high model coefficients in the proposed model. However, in the PARAFAC model, the coefficients for these variables are lower in comparison. The disparity in the estimates of the model coefficients can likely be attributed to the unequal representation of the variability of the three-way array by the two models. As shown in the above findings, the PARAFAC model primarily captures the variability specific to the first group rather than the entirety of the dataset. It implies that the emphasis placed on the first group by the PARAFAC model may result in model coefficients representing this group rather than the entire three-way array. This difference between the two models may result in interpretations and conclusions that are different.

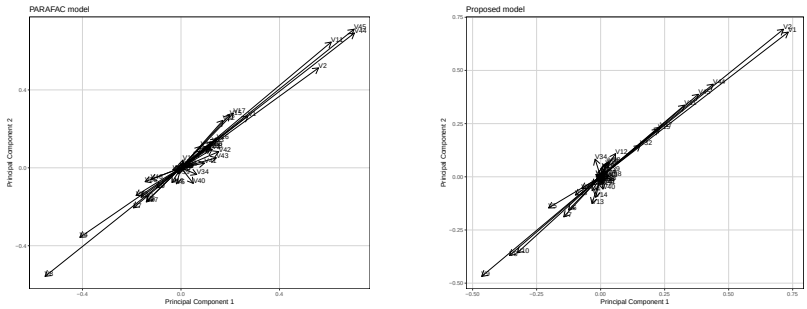


Figure 11: Case study: Biplot of the first two components of the loadings matrix **B** of the PARAFAC and the proposed model.

Following the analysis of the outcomes and consultation with expert engineers in the field, a consensus was reached regarding the results obtained from the proposed model. The variables $V1$ and $V2$ are associated with the pressure applied to each micro-layer, which determines its fusion with the main component. This step is highly important in the production process, as it significantly impacts the numerous mechanical properties of the final product. The variables $V8$ and $V9$ are linked to the bulk temperature, representing the temperature of the main component of all the fused layers. This temperature plays a crucial role in material fusion and adhesion of the individual micro-layers, so the fusion between all the micro-layers influences the printing process overall success and contributes to several quality aspects of the end products. Moreover, the variable $V10$ is linked to the creation of each individual micro-layer, underscoring its significance in the overall process. Lastly, both models highlighted the importance of the variables $V44$ and $V45$, which pertain to the sensors in the cooling step. Cooling temperatures are of noteworthy influence on warpage, a factor directly affecting various qualities of the final product. These findings reinforce the understanding that careful management of cooling temperatures is imperative for ensuring optimal product outcomes across multiple quality dimensions.

Discussion

Across all four data sets, we demonstrated that the proposed approach was able to provide a model that is representative of all the groups present in the data. This was supported by similar levels of variance explained across all groups compared to the variance discrepancies in the PARAFAC model. This is indicative of a model that is representative of the variability of the entire data set and all groups. Despite the fact that one or more groups may exhibit larger variability, the suggested model was able to reduce this variability and produce model coefficients that are common to all groups. The two simulation strategies depicted two ways of inducing a multi-group structure in the data and in both cases, the proposed model was able to regulate the dominance of the group with the highest variance. In the case study, we presented a real-world application in additive manufacturing, aiming to gain valuable process insights and understanding. To achieve this aim, it is necessary to consider a model that accurately represents all groups in the data set. The proposed model showcases a percentage of explained variance similar for all groups, indicating its ability to effectively capture the common variability of the three-way array and its multiple groups. This stands in contrast to the PARAFAC model, which exhibits a greater representation of the first group and its variability. The strength of the proposed model lies in its capability to mitigate the higher variability observed in the first group and successfully recover a common structure that encompasses all groups. An in-depth analysis of the loadings matrix $\mathbf{B}$ provides further insights for process understanding. The results of the PARAFAC model indicate a lack of homogeneity in variability among the groups, leading to different model coefficients. This aspect holds crucial implications for a deeper understanding of the underlying process, as in the PARAFAC case, the coefficients of the variables are more representative of the first group rather than the entire data set. In this context, the proposed model proves to be more suitable and advantageous for three-way arrays that present a multi-group structure. Furthermore, we recommend extending the analysis to include a meticulous examination of residuals, the three-way matrix $\mathcal{E}$, when conducting further studies on differences between various groups. As the model represents the common structure shared among all groups, analyzing the residuals can shed light on the different variations a specific group represents. However, it is essential to note that such an analysis lies beyond the scope of the proposed model in this paper. Instead, it serves as a natural direction for future studies, warranting careful consideration and examination on a case-by-case basis.

4. Conclusions

The extension of the PARAFAC model proposed in this paper makes it possible to analyze complex data that present a multi-grouping structure. This transfer of methodology from the PARAFAC model to the multi-group settings was defined by constraining the objective function to consider the group structure for the model parameters. We provided the corresponding extension of the ALS algorithm to solve the proposed model. This algorithm was set to be also used in the absence of multi-group data as its initialization corresponds to the PARAFAC model.

Three simulation studies and a real case were used to illustrate the capability of the extended PARAFAC method to render a model that explains the common variability of all the groups. In all applications, using the model without a multi-group structure when the groups have large differences resulted in a model being representative of the most dominant group (i.e. the group with the largest variance). In the real case study in additive manufacturing, we

illustrated the impact of considering the grouping structure to analyze the importance of process variables in this type of data. By employing the extended PARAFAC method, we achieved a comprehensive understanding of the entire manufacturing process, avoiding a biased focus on individual groups. This perspective facilitated a more profound analysis of the data and enabled us to extract valuable insights that would have been overlooked in a model dominated by a single group.

Our results demonstrate that the proposed model performs well in capturing the common and explanatory loadings of the data, even when the variances of the groups are unequal, resulting in a more robust and reliable model. The proposed model effectively represents the entire dataset, and this is of utmost importance for gaining valuable insights and understanding of the process. When a model is heavily influenced by one dominant group, the loadings tend to explain more of that particular group's characteristics, neglecting the broader dynamics of the entire process. The proposed method offers a solution to this challenge, ensuring a more balanced representation of the data and fostering deeper process understanding across all groups.

References

[1] M. Spooner, D. Kold, M. Kulahci, Selecting local constraint for alignment of batch process data with dynamic time warping, Chemometrics and Intelligent Laboratory Systems 167 (2017) 161–170.

[2] D. Louwerse, A. K. Smilde, Multivariate statistical process control of batch processes based on three-way models, Chemical Engineering Science 55 (7) (2000) 1225–1235.

[3] M. Ryckewaert, G. Chaix, D. Héran, A. Zgouz, R. Bendoula, Evaluation of a combination of nir micro-spectrometers to predict chemical properties of sugarcane forage using a multi-block approach, Biosystems Engineering 217 (2022) 18–25.

[4] M. Dyrby, M. Petersen, A. K. Whittaker, L. Lambert, L. Nørgaard, R. Bro, S. B. Engelsen, Analysis of lipoproteins using 2d diffusion-edited nmr spectroscopy and multi-way chemometrics, Analytica Chimica Acta 531 (2) (2005) 209–216.

[5] C. M. Rubingh, S. Bijlsma, R. H. Jellema, K. M. Overkamp, M. J. van der Werf, A. K. Smilde, Analyzing longitudinal microbial metabolomics data, Journal of proteome research 8 (9) (2009) 4319–4327.

[6] S. Wold, P. Geladi, K. Esbensen, J. Öhman, Multi-way principal components-and pls-analysis, Journal of chemometrics 1 (1) (1987) 41–56.

[7] R. Bro, Multiway calibration. multilinear pls, Journal of chemometrics 10 (1) (1996) 47–61.

[8] R. Harshman, Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis, UCLA Working Papers in Phonetics 16 (1970).

[9] L. R. Tucker, et al., The extension of factor analysis to three-dimensional matrices, Contributions to mathematical psychology 110119 (1964).

[10] R. Vitale, O. E. de Noord, J. A. Westerhuis, A. K. Smilde, A. Ferrer, Divide et impera: How disentangling common and distinctive variability in multiset data analysis can aid industrial process troubleshooting and understanding, Journal of Chemometrics 35 (2) (2021) e3266.

[11] Y. Ben-Shabat, M. Lindenbaum, A. Fischer, 3dmfv: Three-dimensional point cloud classification in real-time using convolutional neural networks, IEEE Robotics and Automation Letters 3 (4) (2018) 3145–3152.

[12] C. Ma, Y. Guo, Y. Lei, W. An, Binary volumetric convolutional neural networks for 3-d object recognition, IEEE Transactions on Instrumentation and Measurement 68 (1) (2018) 38–48.

[13] J. D. Carroll, J.-J. Chang, Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition, Psychometrika 35 (3) (1970) 283–319.

[14] L. R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (3) (1966) 279–311.

[15] R. Bro, Parafac. tutorial and applications, Chemometrics and intelligent laboratory systems 38 (2) (1997) 149–171.

[16] A. Eslami, A. Kohler, M. El Qannari, S. Bougeard, General overview of methods of analysis of multi-group datasets., in: HDSDA, 2011, pp. 108–123.

[17] S. Legleye, A. Eslami, S. Bougeard, Assessing the structure of the cast (cannabis abuse screening test) in 13 european countries using multigroup analyses, International journal of methods in psychiatric research 26 (1) (2017) e1552.

[18] A. Tenenhaus, M. Tenenhaus, Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis, European Journal of operational research 238 (2) (2014) 391–403.

[19] M. Hanafi, E. M. Qannari, B. Jaillais, Multi-Block and Three-Way Data Analysis, 2019. doi:10.1016/B978-0-12-409547-2.14717-1.

[20] A. Eslami, E. M. Qannari, A. Kohler, S. Bougeard, Algorithms for multi-group pls, Journal of Chemometrics 28 (3) (2014) 192–201.

[21] J. Kallus, P. Johansson, S. Nelander, R. Jörnsten, Mm-pca: integrative analysis of multi-group and multi-view data, arXiv preprint arXiv:1911.04927 (2019).

[22] W. Krzanowski, Between-groups comparison of principal components, Journal of the American Statistical Association 74 (367) (1979) 703–707.

[23] W. Krzanowski, Principal component analysis in the presence of group structure, Journal of the Royal Statistical Society: Series C (Applied Statistics) 33 (2) (1984) 164–168.

[24] B. N. Flury, Common principal components in k groups, Journal of the American Statistical Association 79 (388) (1984) 892–898.

[25] A. Eslami, E. Qannari, A. Kohler, S. Bougeard, Multivariate analysis of multiblock and multigroup data, Chemometrics and Intelligent Laboratory Systems 133 (2014) 63–69.

[26] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (3) (2009) 455–500.

[27] J.-h. Jiang, H.-l. Wu, Y. Li, R.-q. Yu, Three-way data resolution by alternating slice-wise diagonalization (asd) method, Journal of Chemometrics: A Journal of the Chemometrics Society 14 (1) (2000) 15–36.

[28] P. Paatero, A weighted non-negative least squares algorithm for three-way 'parafac'factor analysis, Chemometrics and Intelligent Laboratory Systems 38 (2) (1997) 223–242.

[29] N. Faber, L. Buydens, G. Kateman, Generalized rank annihilation method. i: Derivation of eigenvalue problems, Journal of chemometrics 8 (2) (1994) 147–154.
[30] E. Sanchez, B. R. Kowalski, Tensorial resolution: a direct trilinear decomposition, Journal of Chemometrics 4 (1) (1990) 29–45.
[31] H.-P. Yeh, K.Meinert, M.Bayat, J.Hattel, Part-scale thermo-mechanical modelling for the transfusion module in the selective thermoplastic electrophotographic process, in: WCCM-APCOM 2022, Vol. 1000, Manufacturing and Materials Processing, 2022.
[32] Y. Wu, K. He, Group normalization, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

CHAPTER 6

Multi-group N-PLS

6.1 Introduction

In this section, we present a novel model aimed at the analysis of process data and associated quality data. This method represents a significant extension of the methodology, the multi-group PARAFAC model, as delineated in Chapter 5, in order to consider the integration of quality-related data. The motivation for this model is to provide a more thorough understanding of processes, create opportunities for process improvement and create the opportunity for outcome prediction.

The previous Chapter introduced the multi-group PARAFAC model, an analytical technique designed to unravel intricate patterns inside complex three-way data sets. The model is part of the so-called unsupervised models, meaning that it discerns latent structures and relationships within the data without considering any quality variables. The unsupervised nature of the multi-group PARAFAC model has several advantages in uncovering latent relationships, providing a comprehensive understanding of the multi-way data structure, and finding patterns among groups and various batches. However, there are some limitations associated with the model, especially when the objective is to exploit data insights for the purpose of process optimization.

In order to overcome these limitations and to fully utilize the capabilities of our analytical framework, a progressive transition from an unsupervised to a supervised model was conducted. The introduction of additional data, referred to as the response variable $\mathbf{Y}$, becomes essential in light of this extension. Using this variable enhances our model's capability to identify inherent latent patterns and analyze the impact of $\mathbf{Y}$ on the resulting model.

Therefore, the shift from an unsupervised model to a supervised model represents a fundamental change in the model perspective. This model enables us to build a model and model loadings that are exemplary of the input data but also take into account their interaction with the output variable, thus making the model much more useful for a targeted investigation into optimization and process understanding. By taking into account the response variable $\mathbf{Y}$, we are able to not only understand the interactions between observations, variables, batches and responses but also anticipate their potential predictions.

In this case study, we considered the process data as input data and the quality of the final products as the output data. The input data are organized into three-way

arrays that are represented by three distinct modes, namely modes a, b and c. Each of these modes incorporates distinct information aspects that provide a thorough description of the data acquired from the process being examined. Mode a represents the observations acquired during the course of the production, mode b represents the variables collected throughout the multiple sensors placed in different stages of the production and mode c represents the batches that constitute the iterative instances of the process, refer to Figure 6.1. The link between the different rows of products and the observations collected during their production creates a grouping structure.

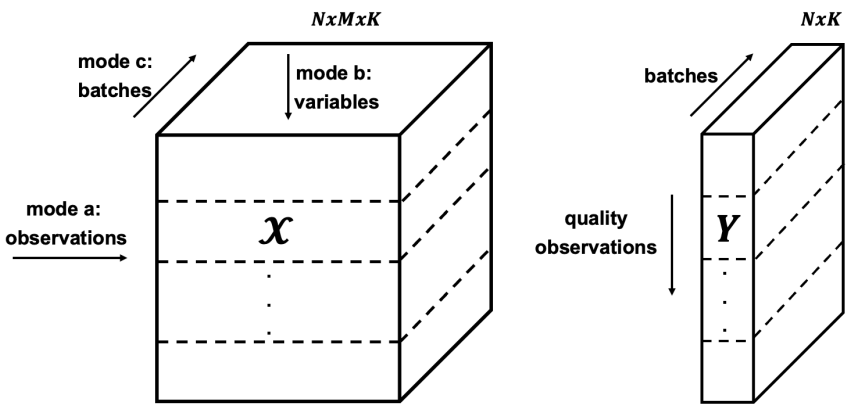


Figure 6.1. Representation of the input and output multi-group data.

The variable $\mathbf{Y}$ can be represented into two-way arrays form with a grouping structure. In the particular case of additive manufacturing, the output variable $\mathbf{Y}$ represents a distinctive characteristic or quality related to each specific batch. The output $\mathbf{Y}$ is represented as a two-way array where columns represent the batches and rows represent the observed quality measurements. 6.1.

The inclusion of variable $\mathbf{Y}$ as a quality column for each batch highlights the complexity of the study, as it encompasses the performance and quality metrics linked to the produced products. Quality data is a fundamental piece of information that augments the comprehensive understanding of the process dynamics. The correlation between the input variables $\mathcal{X}$ and the quality characteristics $\mathbf{Y}$ highlights the relevance of our analytical methodology, especially in the field of additive manufacturing, where maintaining product quality is of utmost importance.

The presentation of the multi-group multi-way Partial Least Squares (multi-group N-PLS) methodology is included in the draft of the paper in the following section. The initial version of the article presents a thorough examination of the multi-group NPLS technique, including a detailed explanation of its theoretical framework and technical aspects. The methodology presented in this study is considered as prelimi-

nary findings. As such, these preliminary results are not intended to be an endpoint but rather represent a foundation for research laying the groundwork for future work, promoting deeper explorations and more comprehensive validations of the multi-group N-PLS methodologies.

6.2 Manuscript: Multi-way PLS for the analysis of multi-group three-way data

Draft Paper: Marta Rotari and Murat Kulahci, ”Multi-way PLS for the analysis of multi-group three-way data”.

Multi-way PLS for the analysis of multi-group three-way data

Marta Rotari[a], Murat Kulahci[a,b]

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

Abstract

This paper introduces a novel approach to analyze multi-group three-way arrays. The introduced method is an extension of the multi-way Partial Least Square method, designed to incorporate the multi-group structure observed in the data. The concept of a multi-group structure arises from the classification of data into multiple groups based on common identifying factors. This type of data is often observed in industrial manufacturing cases where the data are characterized by observations, variables and batches organized in three-way arrays. The association between the final products and observations acquired during their production create the grouping structure. This paper addresses the existing gap in the exploration of supervised three-way array models within multi-group configurations. Conventional methods often overlook intrinsic groupings, leading to a mixture of between-group and within-group variances. Conversely, this study introduces a model that integrates input and response variables while considering the group structure.

Keywords: multi-way N-PLS; multi-group data; three-way data; supervised method

1. Introduction

Modern industrial production systems incorporate sophisticated technologies capable of gathering data from many sources, resulting in an abundance of data. This often leads to data sets organized in multi-group three-dimensional arrays. Three-way arrays are data formats containing observations, variables and conditions that could represent different factors such as time, batches, and other contextual conditions.

As an example, in the context of a batch fermentation process, a collection of time-dependent data that encompasses numerous variables and batches can be represented as a three-dimensional array [1]. Another example is in systems that depend on sensory data, which may originate from several channels and many devices simultaneously [2, 3].

The notion of a multi-group structure refers to the classification of observations into distinct groups. The creation of these groups occurs when observations or samples exhibit specific distinctive characteristics [4, 5, 6]. These features may include demographic factors, qualitative features, geographical location, or other relevant criteria, thus delineating distinct groups.

Numerous approaches have been developed to address three-way arrays, including both unsupervised and supervised models. Among the unsupervised techniques, prominent approaches such as Parafac and Tucker [7, 8] have been developed to unravel the latent structures within the data. Supervised methods like multi-way Partial Least Squares (PLS) [9, 10, 11, 12] have been developed to effectively integrate regression analysis with tensor factorization for the purpose of mapping data into latent spaces with reduced dimensions. The combination of regression and tensor factorization in this combination provides these approaches with the capability of uncovering complex patterns and relationships, hence improving predicted accuracy and interpretability.

N-PLS regression entails a simultaneous decomposition of $\mathcal{X}$ and $\mathbf{Y}$ through latent variables, capturing maximal covariance between the two. The model is utilized in scenarios where the number of variables is higher than the number of observations. The model is also widely used in problems of dimensionality reduction. By extracting a reduced set of latent variables that incorporate essential information from both $\mathcal{X}$ and $\mathbf{Y}$, the model enables effective modelling, prediction and interpretation framework for N-way data.

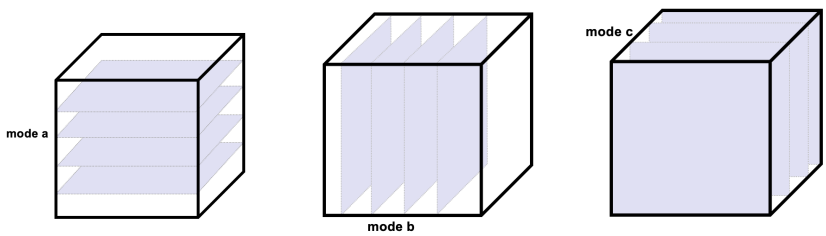
To examine multi-group structured data sets, many models for two-way arrays have been developed [4, 5, 6]. Yet, no analogous extension for supervised three-way array models has been identified thus far. Assuming $\mathcal{X}$ and $\mathbf{Y}$ are

partitioned into G groups, a straightforward approach involves performing an N-PLS regression without acknowledging the group structure. However, this disregards individual groupings and mixes between-group and within-group variances. An alternative approach involves applying N-PLS regression separately to each group. However, this approach leads to analysing too many parameters and, therefore, can hinder the compression of the model or lead to inconsistent and misleading conclusions.

In this paper, we present an extension of the N-PLS model to the case of multi-group three-way data sets. The proposed model aims to construct a model and model parameters adjusted according to the multi-group structure of the input data $\mathcal{X}$ and the response $\mathbf{Y}$. The model aims to identify a common structure by calculating common loadings in the presence of multiple groups in the data sets. This model aims to examine the connections between variables and different experimental situations in relation to the response variable $\mathbf{Y}$ through the exploration of the covariance structure. As a result, the loadings matrices produced by the proposed model capture the importance of individual variables and experimental circumstances in the overall variability seen in the three-way data, representing all constituent groups included in the array. By identifying common loadings, the proposed model provides a more accurate and comprehensive description of the underlying system.

2. Notations

Three-way arrays are denoted by $\mathcal{X} \in \mathbb{R}^{N \times M \times K}$ whose elements are x_{nmk} where $n = 1, \ldots, N$, $m = 1, \ldots, M$ and $k = 1, \ldots, K$. The three-way entries are represented by three different modes a, b and c as shown in Figure 1. Mode a represents the data $\mathcal{X}_{n\bullet\bullet}$ and usually collects the observations. Mode b represents the data $\mathcal{X}_{\bullet m\bullet}$ and is the mode related to the variables. Finally, mode c represents different batches or conditions, such as time, different levels of temperatures and represents the data $\mathcal{X}_{\bullet\bullet k}$, $k \in 1, \ldots, K$.



(a) Mode-a: Observations. (b) Mode-b: Variables. (c) Mode-c: Batches, conditions.

Figure 1: Modes representation and data organization.

Matrices, or two-way arrays in general, are indicated as $\mathbf{W}$ capital bold letters. One-way arrays, commonly called vectors, are indicated by lower case bold letters $\mathbf{w}$. Horizontal vector concatenation to generate a matrix is indicated by $[\ :\ ]$, $\mathbf{W} = [\mathbf{w}_1 : \ldots : \mathbf{w}_F]$. The letters N, M, K are reserved for specifying dimensions, while the letter F is reserved for denoting the number of components. The transposition operator is represented by the super-index ^{T}.

In a multi-group structure, the three-way array $\mathcal{X}$ consists of observations in mode a that are partitioned into G groups. A visual representation of this structure is shown in Figure 2. Each group $g \in \{1, \ldots, G\}$ is represented by $1, \ldots, n_g$ rows, such that $\sum_{g=1}^{G} n_g = N$. The sub-array that contains the observations belonging to group g is represented by $\mathcal{X}_g \in \mathbb{R}^{n_g \times M \times K}$ where n_g is the number of observations in such group. The response multi-group matrix $\mathbf{Y} \in \mathbb{R}^{N \times K}$ contains the response values for N observations and K batches.

2.1. Unfolding

The notion of unfolding or matricising a three-way array has significant importance in the multi-way data analysis domain. Unfolding techniques regard rearranging the data of a three-way array into a matrix, with the specific configuration dependent on the mode under consideration. Mode-a unfolding involves rearranging the three-way array

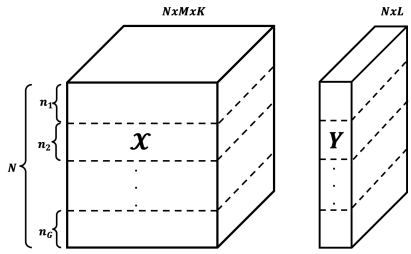


Figure 2: Multi-group representation of the three-way array $\mathcal{X} = x_{ijk} \in \mathbb{R}^{N\times M\times K}$ and $\mathbf{Y} \in \mathbb{R}^{N\times K}$. The observations are divided into G groups, such that group $g = \{1, \cdots, G\}$ is represented by $1, \cdots, n_g$ rows and $\sum_{g=1}^{G} n_g = N$.

$\mathcal{X}$ along the first mode, resulting in a matrix of dimensions $N \times MK$. The mode-a unfolded matrix preserves the multi-group structure, as represented in Figure 3. Mode-b unfolding involves rearranging the three-way array $\mathcal{X}$ along the second mode b (the columns), resulting in a matrix of dimensions $M \times NK$. Finally, mode-c unfolding involves rearranging the three-way array $\mathcal{X}$ along the third mode c, resulting in a matrix of dimensions $K \times NM$.

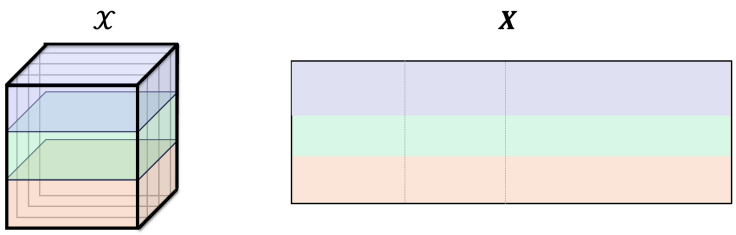


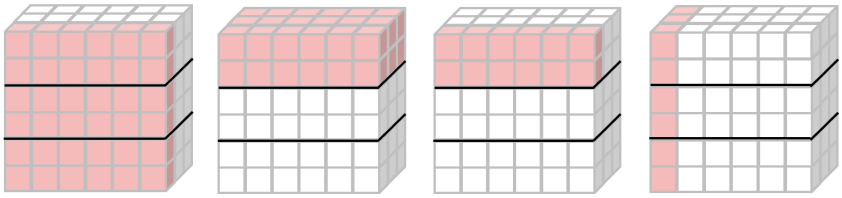
Figure 3: Unfolding of a multi-group three-way array into a multi-group two-way array.

2.2. *Pre-processing*

Scaling is a fundamental data pre-processing technique and represents a fundamental step in data analysis. Standardization and centring are the most often employed scaling strategies in the analysis of three-way data. Standardization is a scaling technique that aims to standardize or transform the data to enhance interpretability and comparability. It is also known as Z-score scaling, which is a specific scaling technique that transforms the data to have a mean of 0 and a standard deviation of 1. Centring involves shifting the three-way or two-way elements to have a zero mean. Usually, these types of pre-precessing are done by columns or by rows, depending on the specific situation.

There are several methods available for scaling a multi-group three-way array[13]. The selection of an appropriate scaling technique depends on several factors, including the inherent characteristics of the data, the specific dimensions to emphasis and the overarching research objectives. Figure 4 illustrates the diverse scaling strategies.

- Batch Scaling: this strategy entails the scaling of data concerning individual batches, denoted as $\mathcal{X}_{\bullet\bullet k}$ for $k \in 1, \ldots, K$.



(a) Batch scaling. (b) Group scaling. (c) Instance scaling. (d) Layer scaling.

Figure 4: Scaling techniques for multi-group three-way arrays.

- Group Scaling: this approach involves scaling data with respect to distinct groups within the three-way array $\mathcal{X}_g \in \mathbb{R}^{n_g \times M \times K}$ for $g \in 1, \ldots, G$. In this instance, greater emphasis is placed on highlighting the differences among the different groupings.
- Instance Scaling: the intersection of batch and group considerations results in instance-based scaling denoted as $\mathcal{X}_{\bullet mk}$ for $m \in 1, \ldots, M, k \in 1, \ldots, K$.
- Layer Scaling: this approach is variable scaling $\mathcal{X}_{\bullet m \bullet}$ for $m \in 1, \ldots, M$.

It is essential to notice that the choice of scaling method may vary depending on the data distribution, the specific algorithm being used, and the aim of the analysis. Proper scaling is a crucial step in the data pre-processing pipeline and can significantly impact the success of subsequent analyses or machine learning models.

3. Methods

This section provides a brief overview of the N-PLS method. Subsequently, the proposed model is presented; it is an extended model considering the multi-group structure of the three-way array $\mathcal{X}$ and the response data $\mathbf{Y}$.

3.1. N-PLS

The Multi-way Partial Least Squares (N-PLS) method is an extension of multiple linear regression and principal component analysis. It addresses the high-dimensional data sets by identifying latent structures that capture the maximum covariance between the three-way data set $\mathcal{X}$ and the response variable $\mathbf{Y}$.

Consider the three-way array $\mathcal{X} \in \mathbb{R}^{N \times M \times K}$ and the matrix $\mathbf{Y} \in \mathbb{R}^{N \times K}$. The N-PLS regression model can be written as follows:

$$\begin{aligned} \mathcal{X} &= \mathbf{T}(\mathbf{W}^K \odot \mathbf{W}^J) + \mathcal{R}_x \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{R}_y \end{aligned} \tag{1}$$

where the decomposition matrices for the three-way array $\mathcal{X}$ are: $\mathbf{T} = (t_{nf}) \in \mathbb{R}^{N \times F}$ is the score matrix, $\mathbf{W}^J = (w^J_{mf}) \in \mathbb{R}^{M \times F}$ and $\mathbf{W}^K = (w^K_{kf}) \in \mathbb{R}^{K \times F}$ are the loadings matrices associated with mode b and c, respectively. The matrix $\mathbf{U} = (u_{nf}) \in \mathbb{R}^{N \times F}$ is the score matrix and $\mathbf{Q} = (q_{kf}) \in \mathbb{R}^{K \times F}$ is the loading matrix for the matrix $\mathbf{Y}$. $\mathcal{R}_x$ denotes the three-way residuals in the decomposition of the three-way array $\mathcal{X}$ and $\mathbf{R}_y$ denotes the residuals in the $\mathbf{Y}$ decomposition. The regression model between $\mathcal{X}$ and $\mathbf{Y}$ can be written as the regression model between the respective scores matrices:

$$\mathbf{U} = \mathbf{T}\mathbf{B} + \mathbf{R}_u \tag{2}$$

The matrix $\mathbf{B}$ denotes the regression coefficients obtained as $\mathbf{B} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{TU}$. Finally, the $\odot$ denotes the Khatri–Rao product defined as the column-wise Kronecker product. The PLS model results from finding the model matrices that maximize the covariance between $\mathcal{X}$ and $\mathbf{Y}$. Further details on the model or the resolution procedure may be found in the following [11, 12].

3.2. Proposed Method

Let us consider the three-way array $\mathcal{X} \in \mathbb{R}^{N\times M\times K}$ and the matrix $\mathbf{Y} \in \mathbb{R}^{N\times K}$. $\mathcal{X}$ and $\mathbf{Y}$ are partitioned into G groups, as shown in Figure 2. For both $\mathcal{X}$ and $\mathbf{Y}$, each group $g \in \{1, \cdots, G\}$ is represented by $1, \cdots, n_g$ rows, such that $\sum_{g=1}^{G} n_g = N$. The sub-array that contains the observations belonging to group g is represented by $\mathcal{X}_g \in \mathbb{R}^{n_g\times M\times K}$ and $\mathbf{Y}_g \in \mathbb{R}^{n_g\times K}$ where n_g is the number of observations in such group. The N-PLS regression model considering F components can be written in equation (1) and the regression model in equation (2).

The proposed model aims to construct a model and model parameters adjusted according to the multi-group structure of the data. The model aims to identify a common structure by calculating loadings that are representative of all groups in the data. The solving algorithm to find the model matrices computes subsequent components. This implies that during the initial stage, one-order loadings vectors and corresponding scores are computed. Prior to the calculation of subsequent components, the data sets $\mathcal{X}$ and $\mathbf{Y}$ are deflated with respect to the previously computed parameters (model scores and loadings).

The solving algorithm outlines constructing a multi-group N-PLS model consisting of F components. This model is designed to extract latent relationships within three-way data sets, where the three-way array is represented as $\mathcal{X}$, and the corresponding response data is denoted as $\mathbf{Y}$ of dimensions $N \times K$. First, the three-way array $\mathcal{X}$ is unfolded by the a-mode unfolding in the matrix $\mathbf{X}$ of dimensions $N \times MK$. The obtained matrix is a multi-group matrix, as shown in Figure 3.

The algorithm starts by calculating the first component. First, the algorithm calculates the vector $\mathbf{u}$ through the Singular Value Decomposition (SVD) of the response data matrix $\mathbf{Y}$. This selects the response column with the greatest variation. The matrix $\mathbf{Z}_g$ is then computed using the sub-matrix $\mathbf{X}_g$ and the sub-vector $\mathbf{u}_g$ for $g \in 1, \ldots, G$. For each group, the SVD is performed on the matrix $\mathbf{Z}_g$. The dominant left and right singular vectors (associated with the largest singular value) are extracted, yielding the vectors $\mathbf{w}_g^J$ and $\mathbf{w}_g^K$. These matrices capture latent structures within the data for each group. After calculating group-specific vectors, a weighted mean is computed to calculate the global loadings $\mathbf{w}_f^J$ and $\mathbf{w}_f^K$ to capture the overarching latent patterns representing all the groups simultaneously.

The next steps involve calculating the scores matrix $\mathbf{t}_f$, the $\mathbf{Y}$ loading vector $\mathbf{q}_f$, and the ultimate scores vector $\mathbf{u}_f$ for the first component, $f = 1$. All the computed vectors are then added to the respective model matrices. To calculate subsequent components, a deflation step is performed on both $\mathbf{X}$ and $\mathbf{Y}$ to remove the contribution of the already-extracted latent component. The outer loop then proceeds to the next component iteration until all desired components have been extracted.

For each component, the algorithm computes the regression coefficient $\mathbf{b}$ using the scores matrix $\mathbf{T}$ and the vector $\mathbf{u}_f$. The algorithm presents a systematic procedure for developing a multi-group N-PLS model involving the iterative extraction of latent components, group-specific analyses, and regression coefficient computation, all designed to take into account the grouping structure of the three-way array $\mathcal{X}$ and the corresponding response data $\mathbf{Y}$. Algorithm 1 describes the entire procedure of the proposed method.

Algorithm 1 Algorithm for multi-group N-PLS model

Input: $\mathcal{X}$, $\mathbf{Y}$, G list of group identification, F number of components
  Unfold the three-way array $\mathcal{X}$ in the matrix $\mathbf{X}$
  Set $\mathbf{u}$ as a column of a $\mathbf{Y}$
  for $f = 1, \ldots, F$ **do**
    while conv or max_iter. **do**
      for $g \in 1, \ldots, G$ **do**
        $\mathbf{Z}_g = \mathbf{u}_g^T \mathbf{X}_g$
        $\mathbf{w}_g^J, \mathbf{w}_g^K = SVD(\mathbf{Z}_g)$
      end for
      $\mathbf{w}_f^J = w_mean(\mathbf{w}_1^J, \ldots, \mathbf{w}_G^J)$
      $\mathbf{w}_f^K = w_mean(\mathbf{w}_1^K, \ldots, \mathbf{w}_G^K)$
      Calculate $\mathbf{t}_f$
      $\mathbf{q}_f = \mathbf{Y}^T \mathbf{t}_f$
      $\mathbf{u}_f = \mathbf{Y}\mathbf{q}_f$
      Check convergence
    end while
    Concatenate the matrices: $\mathbf{T}$, $\mathbf{W^J}$, $\mathbf{W^K}$, $\mathbf{Q}$, $\mathbf{U}$
    $\mathbf{B} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}\mathbf{u}_f$
    Deflate $\mathbf{X}$
    Deflate $\mathbf{Y}$
  end for

Comments. In a simple case where the response $\mathbf{Y}$ is represented by a one-way array or vector of dimensions $N \times 1$, the algorithm is exemplified. The matrix $\mathbf{U}$ becomes the $\mathbf{Y}$ matrix and the matrix $\mathbf{Q}$ is a matrix of ones.

The number of components F to choose is usually evaluated using cross-validation (CV) techniques [14, 15]. In general, the number of components depends upon the data and the specific objective of the instance.

To compute the global loadings $\mathbf{w}_{\bullet f}^J$ and $\mathbf{w}_{\bullet f}^K$ matrices, the algorithm suggests using a weighted mean of the loadings vectors computed for each group. Nevertheless, the determination of the weights may vary depending on the unique circumstances of each instance and the intended purpose of the study. Indeed, the context may encompass divergent scenarios. One such scenario is the desire to obtain a model that simultaneously represents all the groups in the data set. In this scenario, a simple mean of the loadings could be computed or a weighted mean. In the case of the weighted mean, the weights are computed taking into account the explained variance per group calculated at the preceding iteration. This entails that a higher weight is assigned to the group with a lesser explained variance in order to ensure that the loading matrices provide an equal level of explanation for all groups. A second scenario is when a thorough understanding of the collected data is possessed. In this particular circumstance, the distribution of weights may not exhibit a uniform trend but instead conforms to specific indications relevant to the problem.

References

[1] M. Spooner, D. Kold, M. Kulahci, Selecting local constraint for alignment of batch process data with dynamic time warping, Chemometrics and Intelligent Laboratory Systems 167 (2017) 161–170.

[2] M. Ryckewaert, G. Chaix, D. Héran, A. Zgouz, R. Bendoula, Evaluation of a combination of nir micro-spectrometers to predict chemical properties of sugarcane forage using a multi-block approach, Biosystems Engineering 217 (2022) 18–25.

[3] M. Dyrby, M. Petersen, A. K. Whittaker, L. Lambert, L. Nørgaard, R. Bro, S. B. Engelsen, Analysis of lipoproteins using 2d diffusion-edited nmr spectroscopy and multi-way chemometrics, Analytica Chimica Acta 531 (2) (2005) 209–216.

[4] A. Eslami, E. M. Qannari, A. Kohler, S. Bougeard, Algorithms for multi-group pls, Journal of Chemometrics 28 (3) (2014) 192–201.

[5] A. Eslami, E. M. Qannari, A. Kohler, S. Bougeard, Multi-group pls regression: application to epidemiology, in: New Perspectives in Partial Least Squares and Related Methods, Springer, 2013, pp. 243–255.

[6] L. Luo, X. Peng, C. Tong, A multigroup framework for fault detection and diagnosis in large-scale multivariate systems, Journal of Process Control 100 (2021) 65–79.

[7] R. Harshman, Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis, UCLA Working Papers in Phonetics 16 (1970).

[8] L. R. Tucker, et al., The extension of factor analysis to three-dimensional matrices, Contributions to mathematical psychology 110119 (1964).
[9] S. Wold, P. Geladi, K. Esbensen, J. Öhman, Multi-way principal components-and pls-analysis, Journal of chemometrics 1 (1) (1987) 41–56.
[10] M. Hanafi, S. S. Ouertani, J. Boccard, G. Mazerolles, S. Rudaz, Multi-way pls regression: Monotony convergence of tri-linear pls2 and optimality of parameters, Computational Statistics & Data Analysis 83 (2015) 129–139.
[11] R. Bro, Multiway calibration. multilinear pls, Journal of chemometrics 10 (1) (1996) 47–61.
[12] S. de Jong, Regression coefficients in multilinear pls, Journal of Chemometrics: A Journal of the Chemometrics Society 12 (1) (1998) 77–81.
[13] Y. Wu, K. He, Group normalization, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[14] E. Kemsley, Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods, Chemometrics and intelligent laboratory systems 33 (1) (1996) 47–61.
[15] H. A. Kiers, Towards a standardized notation and terminology in multiway analysis, Journal of Chemometrics: A Journal of the Chemometrics Society 14 (3) (2000) 105–122.

CHAPTER 7
Collaborative research

7.1 Introduction

This Chapter provides an overview of the collaborative research conducted within the broader initiative of which this PhD is a part project. As discussed in Chapter 1.1, this doctoral project is an integral component of a larger initiative within the MADE organization and entails collaboration with three doctoral students. Part of the project involves mutual collaboration between all PhD students, fostering synergistic cooperation to tackle intricate challenges that extend beyond the boundaries of individual study scopes.

The collaboration relies on the common contribution to tackle difficult problems that are not possible to solve in a single PhD project but require multidisciplinary collaboration. In this case, each PhD student contributes to the development of the problem using their own area of expertise. An example can be in instances where a student, during the course of their investigation, encounters a challenging problem that cannot be solved within the context of their specialized domain. In such cases, such a problem can be presented within the collaborative meetings. These meetings are held involving PhD students and supervisors to discuss the progress of the various projects, receive feedback and comments and decide the next steps. Through this collective engagement, the group endeavours to unravel the problem employing insights from varied fields, such as physics or data analysis and the analysis of the application of novel sensor technology. This multidisciplinary analysis facilitates a new perspective that generates novel strategies and innovative pathways to solve the problem at hand.

Two such collaborations are presented in this chapter, highlighting the collaborative multidisciplinary approach. First is presented the integration of a laser system for layer-by-layer monitoring. The second project aims to study the dimensional defects encountered in the final products. Finally, some of the latest analysis following the dimensional defects problem is presented.

7.2 Laser profiling system

7.2.1 Introduction

This first collaborative project regards a laser profile system integrated within the framework of STEP technology. This project centres around the deployment of a laser profiler consisting of a configuration of four lasers. Its core function is to meticulously assess the stability of the manufacturing process with respect to each incrementally added micro-layer, thus contributing to the production of a three-dimensional bulk object. This assessment is achieved through the acquisition of image data depicting the height of the 3D bulk subsequent to each layer deposition.

The main objective of this project is to ensure the uniformity of each individual micro-layer, necessitating an even distribution of material throughout the process. Achieving uniform products depend on the uniformity of these micro-layers. To this end, the acquired images are scrutinized for potential irregularities, such as holes and bumps, which might compromise the integrity of the final products. Consequently, this technology monitors the production process and assesses the sufficiency or insufficiency of material utilization.

Interconnected with this aim is the introduction of ”fingerprint” test structures within the manufacturing process. These test structures serve as evaluative benchmarks for process performance. Subsequent to the build job, an image detection post-processing methodology is employed to objectively assess the dynamic dimensional conformity of each layer of the manufactured component layer by layer. This analysis provides valuable perspectives on the complexities of the additive manufacturing process and its capacity for achieving precise dimensional accuracy.

Through the application of the real-time data obtained from this quality control procedure, it is possible to achieve a greater level of control over the production process. A consequential feedback loop could be established, facilitating the dynamic adjustment of manufacturing parameters in real-time. This adaptability is crucial in order to optimize the production process and enhance product quality.

Moreover, the outcomes of this collective effort extend the local domain of production. The utilization of these innovative sensors and algorithms in this quality control process could be effectively integrated into the digital twin architecture. This strategic integration enhances the efficacy of the digital twin and augments the technology's applicability across a diverse spectrum of applications.

This project is still ongoing. The initial outcome obtained from the utilization of this innovative technology is collected in the peer-reviewed conference paper presented in the next section.

7.2.2 Paper: In-process monitoring of selective thermoplastic electrophotographic process by laser profiling system and digital fingerprint

S. Shan, H.-P. Yeh, M. Rotari, K. Ælkær Meinert, J. H. Hattel, D. B. Pedersen, M. Kulahci, H. N. Hansen, Y. Zhang, and M. Calaon, "In-process monitoring of selective thermoplastic electrophotographic process by laser profiling system and digital fingerprint", in Proceedings of the 23rd international conference of the european society for precision engineering and nanotechnology, pp. 181–182, euspen Press, May 2023.

euspen's 23rd International Conference & Exhibition, Copenhagen, DK, June 2023
www.euspen.eu

In-process monitoring of selective thermoplastic electrophotographic process by laser profiling system and digital fingerprint

Shuo Shan[1], Hao-Ping Yeh[1], Marta Rotari[2], Kenneth Ælkær Meinert[1], Jesper Henri Hattel[1], David Bue Pedersen[1], Murat Kulahci[2], Hans Nørgaard Hansen[1], Yang Zhang[1], Matteo Calaon[1]

[1]*Department of Mechanical Engineering, Technical University of Denmark, Building 427A, Produktionstorvet, 2800 Kgs. Lyngby, Denmark*
[2]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Building 324, Richard Petersens Plads, 2800 Kgs. Lyngby, Denmark*

sshan@dtu.dk

Abstract

In the past decade, the demand for high-volume production of high-precision products with complex shapes has led to the development of new manufacturing processes. Among these processes, the Selective Thermoplastic Electrophotographic Process (STEP) shows potential for meeting these production demands. STEP involves layer-wise construction of a 3D object from a CAD model. However, during process characterization and optimization, defects such as dimensional flaws have been detected in the final products. To address this quality control issue, a laser profiling system has been installed on the STEP machine to detect process stability for every layer added to the 3D bulk. Additionally, finger print test structures have been introduced in the build job to evaluate process performance. Image detection post-processing of the test geometries is carried out to quantify the dynamic dimensional conformance layer by layer of the manufactured component. By using the online data collected through this quality control process, greater control over the manufacturing process can be achieved, and a feedback loop can be established. This feedback loop will allow for the online adjustment of manufacturing parameters, which will be applicable and provide great help for product optimization. The new sensors and algorithms used for this quality control process will also be incorporated into the digital twin, making this technology even more valuable for a wide range of applications.

Process sensing technologies, 3D scanning, Additive manufacturing, Object detection

1. Introduction

Selective Thermoplastic Electrophotographic Process (STEP) is a new technology in the additive manufacturing field. Electrophotographic imaging [1] enables high-volume production by fusing 2D layers into a single 3D bulk structure. It is composed of two essential modules: an electrophotographic engine and a transfusion module. Unlike other conventional additive manufacturing techniques, STEP has been validated as a potential alternative for the massive production of polymer components thanks to its high speed, good precession and fast starting time [2].

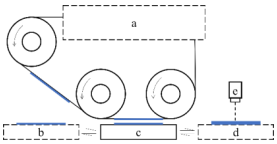


Figure 1. STEP process. a. Electrophotographic Process; b Heating; c. Transfer; d. Cooling; e. Laser scanning

During the STEP printing process, the building plate moves back and forth during the coming layer stacking on it, which makes it difficult to realize precise alignment between successive layers. To address this issue, this investigation proposed a method using digital fingerprint and in-process monitoring to obtain the quality of the printing layer, which will be sent to the printer as feedback. Fingerprints are designed, including the dimensional and positional accuracy of 36 squares as indicators of printing quality. A laser profiling system scans the printed surface and generates the heightmaps. Afterwards, the images are post-processed by transformation and alignment. Object detection and measurement are then applied to the heightmaps to obtain the positions and dimensions of the reference fingerprint.

2. Methodology

2.1. STEP printing and laser profilers alignment

The STEP process is illustrated in Figure 1. The building platform (BP) moves back and forth in 3 positions: heating, transferring and cooling. The BP and the previously printed layers are heated up to 120 degrees in the heating position. When the BP moves to the transferring position, the toner on the drum is then transferred to the BP by melting and fusing. The BP is then cooled down to the cooling position.

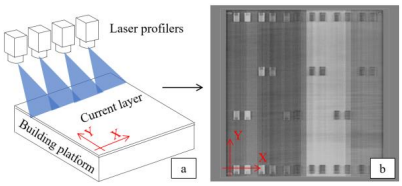


Figure 2. Laser alignment (a) and the generated heightmap of current layer (b).

Figure 2 shows the alignment of four laser profilers that scan the entire building platform. The relative height of the current layer is measured by sampling points on lines. As the BP moves, the sampled lines are stitched and generate a heightmap.

2.2. Finger print object detection and measurment

On the heightmap, 36 finger print structures like squares are designed as references distinguished by the Z direction difference. For each layer, the squares are to be detected with bounding boxes, which are listed as: $[X\ Y\ W\ H\ P]$. X and Y are the absolute coordinates on the heightmap, representing the position of the detected squares; W and H are the width and length of the bounding boxes, as depicted in Figure 3. Finally, P is the confidence of the detection, which is derived based on the mean average precision at the intersection over union, representing the probability of correctly detecting the bounding box.

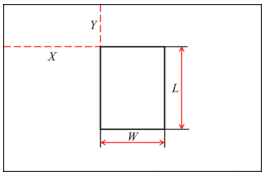


Figure 3. Position and dimension information of the finger print reference squares.

For object detection, a neural network is trained. Data augmentation techniques on images like flipping, rotation, scaling and colour manipulations are performed during the training process. Layered heightmaps derived from the STEP process are submitted to the trained neural network, where positions and dimensions of reference squares are inferenced. The results are then compared to the input mask and the feedback is sent back to the STEP machine to make adjustments. The overall in-process monitoring framework is illustrated in Figure 4.

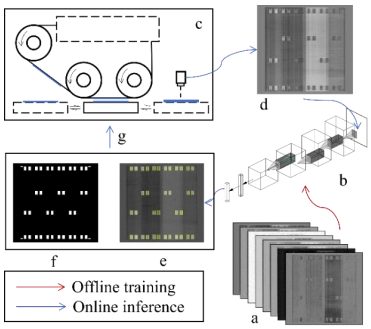


Figure 4. In-process monitoring loop. a. labelled dataset; b. trained neural network; c. STEP process; d. heightmap during the printing process; e. results of reference squares detection; f. input mask; g. feedback to STEP.

3. Results and analysis

The neural network training is conducted on Tesla V100. The training set consists of 30 heightmaps, including 1080 labelled squares, whereas the test set comprises 24 heightmaps. On 21 heightmaps, the model correctly identified all the squares, whereas in the rest 3 heightmaps, there are 34, 22, and 17 detected, respectively.

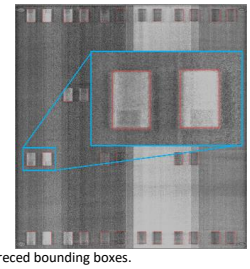


Figure 5. Infereced bounding boxes.

Figure 5 shows the inference bounding boxes of the reference squares. It can be noticed that the reference finger print squares are well detected with dimensions bordered by bounding boxes. The inferencing process takes less than 0.6 seconds on the CPU (Intel(R) Core(TM) i7-1185G7 @ 3.00GHz). Considering the printing time for each layer, the proposed approach is sufficient for in-process monitoring, online printing quality inspection and corrective feedback loop for the STEP process. The confidences of inference squares on corners are shown in Figure 6. The average confidence ranges from 85.2% to 93.3%, with up to 5.6% deviation.

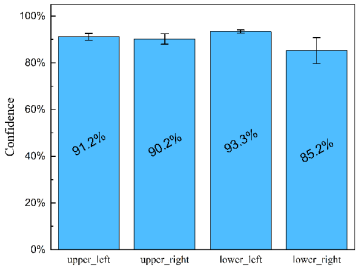


Figure 6. Confidences of inferences of corner reference squares.

4. Conclusion and future works

This investigation proposed an in-process quality inspection and monitoring approach for the STEP process. A laser profiler system is employed to characterize the current printed layer as a heightmap, which is then sent to a neural network to obtain the positions and dimensions of finger print reference squares manufactured as an integral part of the produced component. The results of the experiments show good performance speed. The approach proved a good fit for building a close sensoring loop in the STEP process. Future works include using input masks to crop the heightmap so that the inference areas can be greatly decreased, thus the inference speed can be boosted and other geometries other than reference squares can be applied in the approach.

References

[1] Kumar A V, Dutta A and Fay J E 2004 *Rapid Prototyping Journal.*

[2] Hanson W J, Sanders J R, Bacus M W and Chillscyzn S A 2011 *Electrophotography-based additive manufacturing system with transfer-medium service loops.*

7.3 Dimensional defects

7.3.1 Introduction

The second joint project focuses on the analysis of dimensional faults observed in the final products. The optimisation of manufacturing processes is mainly focused on enhancing the quality of the end products and the dimensions of the final products are among the features that are taken into consideration. Following the completion of the production process, the items undergo a thorough examination and several measurements are recorded for all the end products. The acquired measurements are then thoroughly examined to determine their conformity with the precise parameters stated in the initial 3D (CAD) digital model.

At the basis of this project lies the development of a model based on physical principles to describe the process of heat transfer. The presented model represents a notable development in understanding the complexities that regulate heat transfer phenomena, specifically within the framework of the STEP technology. The effectiveness of this model becomes evident through its comparison with the empirical results obtained from the end products produced by the STEP technology. This comparative analysis validates the model's ability to accurately anticipate real-world manufacturing process outcomes. The alignment or deviation between the model predictions and the actual product measurements provides valuable insights into the interaction between heat transport processes and dimensional results.

Furthermore, it is important to note that this project is still undergoing active investigation of dimensional defects. By consistently prioritising product quality, the collaborative effort shows a firm commitment to systematically investigating and comprehending the various factors that contribute to dimensional deviations. The initial outcomes are collected in the peer-reviewed conference paper presented in the next section.

7.3.2 Paper: Thermo-mechanical model for a selective thermoplastic electrophotographic process for dimensional defects

H.-P. Yeh, M. Rotari, S. Shan, K. Ælkær Meinert, J. H. Hattel, M. Kulahci, D. B. Pedersen, and M. Calaon, "Thermo-mechanical model for a selective thermoplastic electrophotographic process for dimensional defects", in Proceedings of the 23rd international conference of the european society for precision engineering and nanotechnology, pp. 187–188, euspen Press, May 2023.

euspen's 23rd International Conference & Exhibition, Copenhagen, DK, June 2023
www.euspen.eu

Thermo-mechanical model for a selective thermoplastic electrophotographic process for dimensional defects

Hao-Ping Yeh[1], Marta Rotari[2], Shuo Shan[1], Kenneth Ælkær Meinert[1], Jesper Henri Hattel[1], Murat Kulahci[2,3], David Bue Pedersen[1], Matteo Calaon[1]

[1] *Department of Civil and Mechanical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark*
[2]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark*
[3]*Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden*

haoye@dtu.dk

Abstract
Additive manufacturing is a revolutionary method that relieves industries from the geometrical restrictions of their products. In this area, the Selective thermoplastic electrophotographic process (STEP) is a breakthrough approach that can obtain fast and high-volume production. Despite the STEP's excellent characteristics and achievements, the final products present dimensional defects which significantly impact their mechanical qualities. This study presents a prediction model of deformation behaviour by a finite element-based thermo-mechanical model. Additionally, a dimensional study was conducted on various STEP manufactures products. The work presents an analysis of the results of the dimensional measurements, highlighting the position of the products on the building plate. The dimensional evaluation of the STEP's products supports the thermo-mechanical model results.

Additive manufacturing, thermo-mechanical model, dimensional defects, selective thermoplastic electrophotographic process

1. Introduction

Selective thermoplastic electrophotographic process (STEP) is a novel additive manufacturing process. It is a brand-new polymer-based additive manufacturing method introduced by Evolve Additive Solutions, Inc. This new technology might be invaluable addition to injection moulding production by producing completely dense, multi-material. The STEP technology is mainly composed of two fundamental modules, electrophotographic and transfusion modules. The manufacture occurs by fusing 2D layers created by the electrophotographic module into a 3D structure. STEP's 2D-to-3D deposition method involves heating both the incoming 2D layer and the component's structure. Next, the incoming 2D layer is fused to the final component using pressure. Finally, the transfusion module is used for deposition.

Some research in the STEP machine for process digitalization, such as multiphysics modelling, data analysis, sensorics collecting and fingerprints, supports its exploration, comprehension and optimization. Based on the knowledge, the finished product was determined to have dimensional faults that may be connected to the transfusion module of the system and negatively affect its mechanical properties.

This paper proposes a thermomechanical model using finite elements that predicts the behaviour of deformation [1], [2], [3]. The model incorporates multi-material samples consisting of component and support material, bringing the model closer to the actual printing process. Moreover, the model simulations are calibrated using measurement data in order to provide an accurate depiction of defects. Using a 3D scanner, the three dimensions of the printed products obtained from the STEP process are measured and recorded. A brief statistical analysis of these results will be provided, highlighting various defects along the three dimensions.

2. Methodology

2.1. Simulation

In this work, the finite element method (FEM) is used to simulate the deformation of printing samples shown in Figure 1. Based on the knowledge of the STEP machine, the defect is induced by the pressure and heat in the transfusion module could be concluded. The simulation follows the actual manufacturing sequence, i.e., idle, pre-heating, transfusion, and cooling. Afterward, the printed build is detached from the machine and cooled to room temperature.

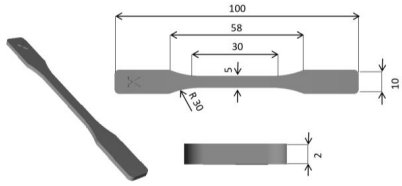


Figure 1. Dimensions of ISO 527 type 1BA sample

The model consists of heat transfer and mechanics theories. Regarding the thermal part, heat conduction, convection, and radiation shown in Equation (1) and Equation (2) are taken into account.

$$\rho c_p \frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right) - \rho \Delta H_{met}\frac{\partial r_{liq}}{\partial T} + \dot{Q}''' \quad (1)$$

It's worthy mention that $\rho \Delta H_{met}\frac{\partial r_{liq}}{\partial T}$ represents the energy caused by melting and solidification.

$$-k\frac{\partial T}{\partial z} = h_{amb}[T - T_{amb}] + \varepsilon\eta[T^4 - T_{amb}^4] \quad (2)$$

Equation (3) of the mechanical model uses Hooke's general rule to calculate stress, strain, and deformation.

$$\sigma_{ij} = \frac{E}{1+v}\left[\frac{1}{2}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) + \frac{v}{1-2v}\delta_{ij}\delta_{kl}\right] e_{kl}^{el} \quad (3)$$

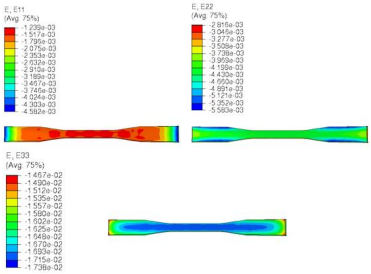


Figure 2. Simulated strain in three directions

For the thermal model, the simulation was validated with the thermal cameras mounted on the machine and the predictive thermal profile was consistent with the sensorics. The printed tensile test samples were used for the validation of the mechanical part of the model, see Section 3.

2.2. Sample measurement data

To investigate the dimensional flaw, the study considered eight batches of seven products each for a total of 56 products. Subsequently, the products were cleaned by removing the support material and measured using a 3D scanner, ATOS ScanBox.

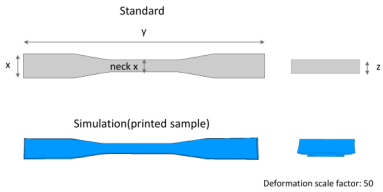


Figure 3. Standard sample vs. printed sample by simulation

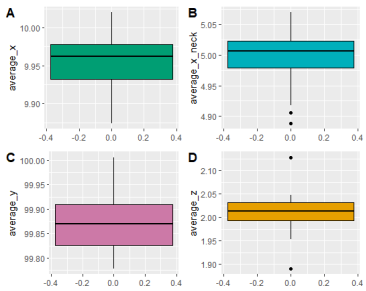


Figure 4. Measurement result for all the dimensions of the STEP's produced products. Figure A: boxplot related to the measurements along the x axis. Figure B: boxplot of the measurements along the $neck\ x$ axis. Figure C: boxplot of the measurements along the y axis. Figure D: boxplot of the measurements along the z axis.

Multiple measurements were collected along the x, y and z axes, as depicted in Figure 3. One measurement was taken along the y-axis, 18 measurements were taken at various positions along the z-axis, and 15 measurements were taken along the x-axis, including 5 for the neck area. In Figure 4 are collected the measurement of the dimensions along the three axes of the printed products. Consequently, these measurements are compared to the one of the thermo-mechanical models.

3. Result and discussion

Based on the process thermo-mechanical model, we get the dimensions of the printed sample and compare them with the original CAD file. The 3-dimensional defects could be calculated as shown in Figure 2. In Table 1 are collected the original measurements, model results and STEP's products analysis results. In order to make the dimensions defects more visible, we apply scale factor 50 to the deformation. Furthermore, Figure 2 shows the nonuniform strain distribution. It indicates that two ends of the sample have higher contraction ratios in the x and y directions. On the other hand, the middle part of sample contracts more in the z direction than the others.

Table 1. Dimensions of sample for ISO standard, simulation, and measurement

Dimensions	Nominal (mm)	Simulation (mm)	Measurement mean (mm)	Measurement std dev (mm)
x	10	9.96	9.95	0.0387
neck x	5	4.97	4.998	0.0388
y	100	99.82	99.87	0.0578
z	2	1.97	2.01	0.0441

According to Table 1, the simulated printed dimensions closely match the average STEP's measurement result. Based on the simulation, the dimensions defects are because of the thermal-induced residual stress. Besides, the residual stress may cause other defects as well, e.g., warpage. The connection between manufacturing parameters and the quality of products could be obtained and the study also provides an opportunity for product optimization to fit the desired tolerance 10 to 20 microns. Therefore, the optimal temperature of the process will be our next goal.

4. Conclusion

The paper suggests a thermo-mechanical numerical model for the STEP transfusion module that is preliminary verified using part quality measurement to better examine the procedure and enhance the quality of the final result. In this research, a thorough manufacturing parametric investigation is also offered. With the help of this modelling strategy, a reliable digital twin of the STEP process could be created and afterwards incorporated into the STEP process itself, acting as a source of feedback for real-time adjustment of the input process parameters to produce a defect-free final product.

References

[1] M. Bayat et al., "Part-scale thermo-mechanical modelling of distortions in Laser Powder Bed Fusion – Analysis of the sequential flash heating method with experimental validation," Additive Manufacturing, vol. 36, Dec. 2020, doi: 10.1016/j.addma.2020.101508.

[2] A. el Moumen, M. Tarfaoui, and K. Lafdi, "Modelling of the temperature and residual stress fields during 3D printing of polymer composites," The International Journal of Advanced Manufacturing Technology, vol. 104, no. 5, pp. 1661–1676, 2019.

[3] A. el Moumen, M. Tarfaoui, and K. Lafdi, "Modelling of the temperature and residual stress fields during 3D printing of polymer composites," The International Journal of Advanced Manufacturing Technology, vol. 104, no. 5, pp. 1661–1676, 2019.

7.4 Further analysis

In continuation of the research described in Section 7.3, concerning the dimensional defects, subsequent investigation and analysis have been done. Central to this study is examining the dimensions of the final products along the three axes in relation to their position on the building plate. This involves the careful measurement and documentation of the dimensions for each completed product. This phase centred on exploring how the position of products on the building plate influences the final product dimensions. Various batches of seven ISO 527 tensile bars were obtained, each batch obtained with a distinct layout or arrangement of the bars. Some examples can be seen in Figure 7.1.

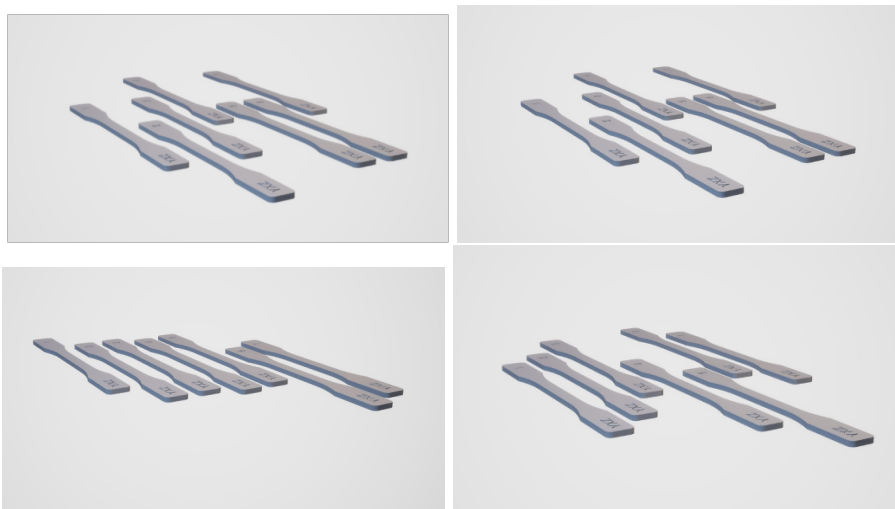


Figure 7.1. Different layout.

Following the completion of the manufacturing process and the realization of the final tensor bars, the meteorological evaluation was conducted along the x, y and z axes, covering the whole three-dimensional space. Refer to Figure 7.2 for visualization. The study considered 18 batches of seven tensile bars each for a total of 126 bars. Subsequently, the bars were cleaned by removing the support material and measured using a 3D scanner, ATOS ScanBox. Multiple measurements were collected along the three axes.

The primary objective of this investigative phase is to identify and define any observable patterns, trends, or inconsistencies that might be attributed to the layout of bars inside the building plate. The study aims to examine the relationship between product positioning and dimensional measurements. The aim is to provide

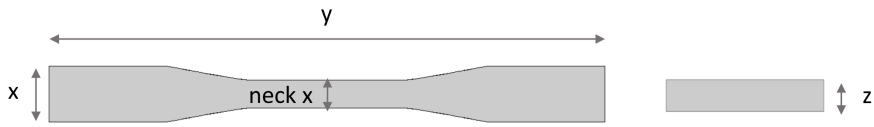


Figure 7.2. Product measurement axes.

substantive insights that contribute to the advancement of theoretical and practical implantation of models and algorithms. The findings of this study aim to enhance our understanding of how the precise placement of the products, together with any interactions between products, can lead to differences in their dimensional qualities.

The measurements obtained were analyzed utilizing several statistical models, resulting in the subsequent findings. No specific patterns or trends were identified along the $y-$axis in connection with the layout or the position of the bars, refer to Figure 7.3. Hence, we concluded that factors such as product location and layout are not factors that influence dimensional faults in the y dimension. However, a different scenario was observed in the x and z dimensions. In these dimensions, the analysis highlighted the presence of defects and discernible patterns. The observations made in this analysis provide valuable insights into the nature of defects and variations in product dimensions, specifically emphasizing the areas where improvements or corrective measures might be needed.

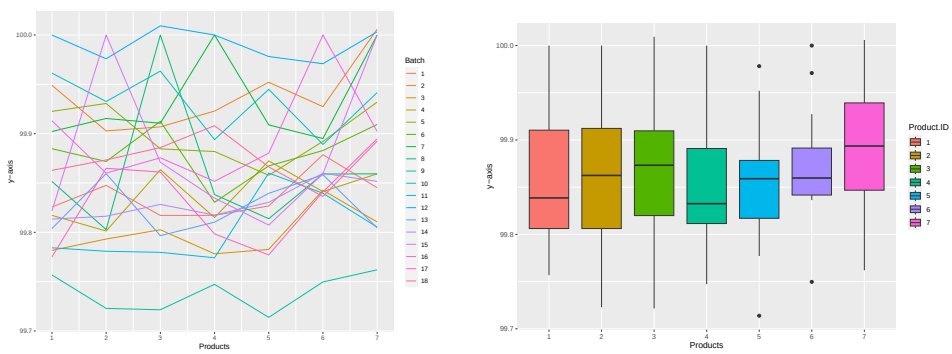


Figure 7.3. Dimensions assessed along the y-axis.

Dimensional defects z-axis The analysis of the z-axis has revealed the presence of discernible irregularities and patterns, refer to Figure 7.4. These findings show that

the thickness of the bars is dependent upon their respective positions on the building plate.

As seen in Figure 7.4, there is a correlation between the thickness of the different bars and their positions within the three-dimensional bulk. This highlights the importance of spatial allocation in determining the geometric features of the products. Evident from Figure 7.4, bars situated at initial positions exhibit a thickness that falls below the established criterion of 2 mm. On the other hand, the bars that are positioned towards the end exhibit a thickness that exceeds the specified value outlined in the 3D CAD file. This phenomenon transcends the diverse batch layouts and demonstrates a systematic correlation with the numerical ordering of product placements.

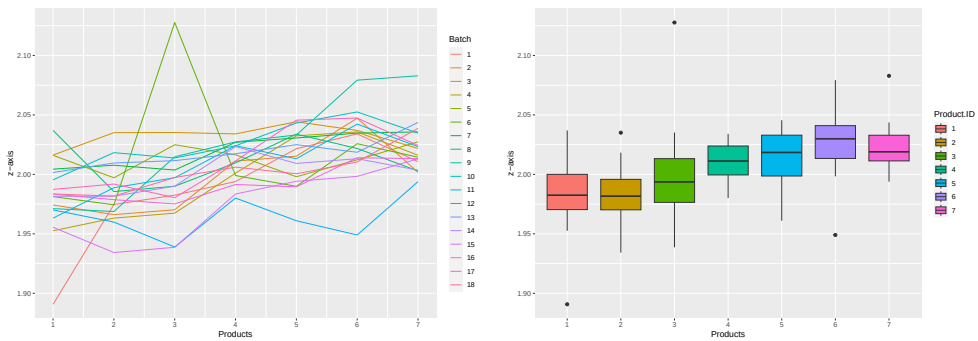


Figure 7.4. Dimensions assessed along the z-axis.

A thorough investigation was conducted by the research team, aiming to explore the potential causal factors underlying the identified defects. A comprehensive range of variables was considered, taking into account aspects such as the thickness of the belt used to transport the micro-layers for deposition onto the three-dimensional bulk. Furthermore, the examination of the equilibrium of pressure applied by the roller during the pressure process of each subsequent micro-layer onto the existing three-dimensional bulk has emerged as a potential factor that deserves further investigation.

After various analyzes, particularly indications pointed towards the possibility of an unbalance in the pressure distribution exerted by the roller during the micro-layer compaction process. This hypothesis was substantiated through the empirical evidence delineated in Figure 7.5. The Figure illustrates the settled pressure point for the pressure in black, the pressure readings obtained from the two sensors positioned at the opposite edges of the roller are represented by red and blue. The lack of alignment between the two readings indicates a disparity in the pressure levels.

The observed disparity not only provides support for the hypothesized imbalance in pressure but also suggests that the forces exerted by the roller are asymmetrical,

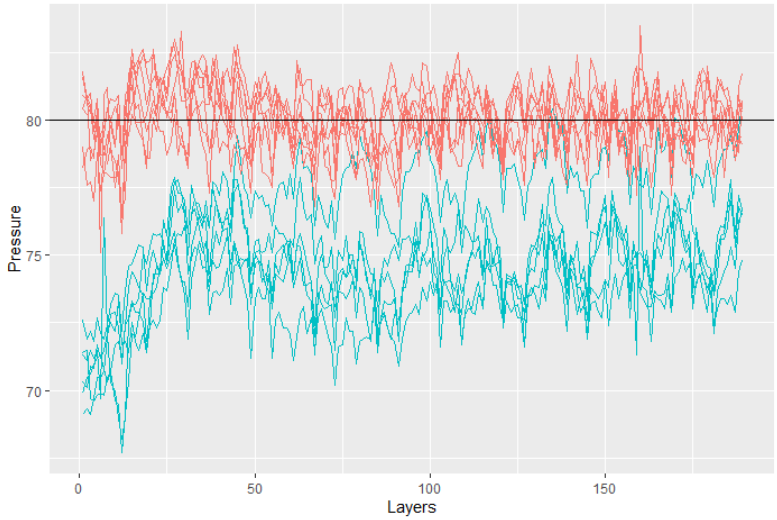


Figure 7.5. Imbalance in the pressure

which might potentially contribute to the occurrence of the reported flaws. Nevertheless, the inferred correlation between the detected disparity and the resultant defects of the bars recorded along the z-axis necessitates further validation. Additional in-depth examinations are indispensable to authenticate this hypothesis. Rigorous and methodical tests and assessments are thus mandated to ascertain the causal linkage between the detected roller pressure imbalance and the discerned dimensional irregularities.

Dimensional defects along the x-axis The investigation has revealed the presence of discernible defects along the $x-$axis, as can be seen in Figure 7.6. Similar to the trends observed along the z-axis, the dimensions of the bars within the $x-$axis are also dependent upon their respective positions within the configuration. This phenomenon unveils a distinct pattern characterized by an "up-and-down" trend. This trend signifies that products occupying odd positions manifest dimensions larger than those occupying even positions.

A series of meetings and discussions have taken place in response to the identified patterns and their likely underlying causes. These sessions aimed to provide a platform for the collaborative sharing of multidisciplinary views and opinions with the goal of understanding the underlying causal processes that explain the observed patterns. The discussion involved the examination of many potential causes that might potentially contribute to these defects. Some plausible hypotheses included imbalances in the distribution of heat across the layers during the manufacturing process, latent material irregularities, and other potentially influential variables that might contribute to dimensionality fluctuations have been suggested. Nevertheless,

despite these hypotheses, the exact underlying factor remains unclear, necessitating the pursuit of more investigation paths.

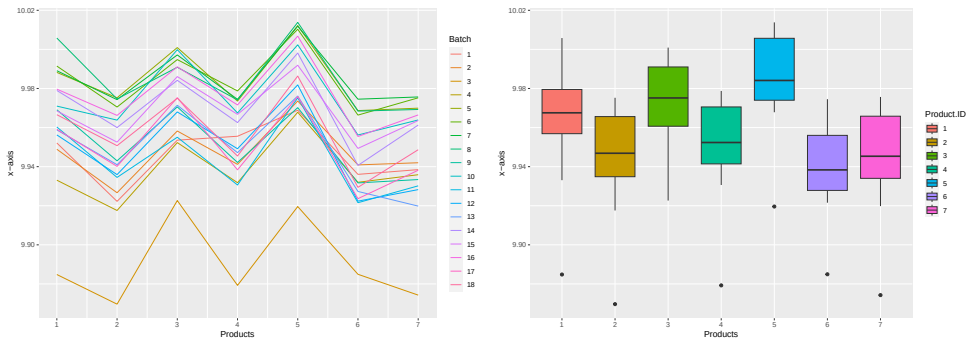


Figure 7.6. Dimensions assessed along the x-axis.

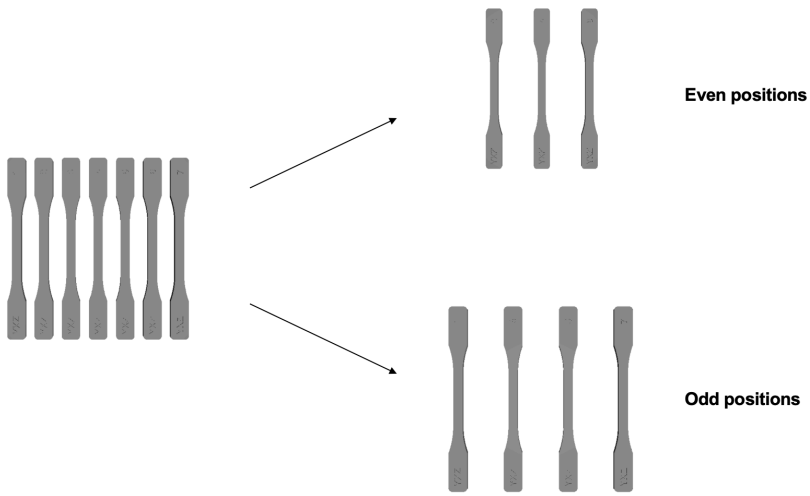


Figure 7.7. Experiments of odd and even positions.

Given the complexity and the elusive nature of the underlying cause, a decisive course of action was established. In the absence of a definitive hypothesis regarding the origins of the observed phenomenon, the research direction was delineated to encompass the execution of two distinct experiments. Still in the stages of ongoing development, these experiments aim to provide significant insights into the complex mechanisms involved. The experimental methodology involves the deliberate categorization and printing of two distinct sets of items: the first involving exclusively

products located in even positions, and the second exclusively featuring products situated in odd positions, refer to Figure 7.7. This approach's conceptual foundation is to isolate and elucidate the differential behaviour between these two subgroups, therefore providing a method to analyse potential causal variables.

CHAPTER 8

Conclusions and future work

This PhD project is part of a large initiative embedded within the MADE organizational framework. The primary objective of this project was to investigate data-driven methodologies for understanding and optimizing additive manufacturing processes, specifically focusing on two prominent technologies: Selective Thermoplastic Electrophotographic Process (STEP) and Selective Laser Sintering (SLS). These processes are advanced and cutting-edge technologies, however, their effectiveness has been limited by an incomplete comprehension of the underlying mechanisms and a substantial requirement for further improvement and optimization. In order to effectively address these limitations, extensive research has been carried out, resulting in the development of a comprehensive range of models. These models have been specifically calibrated to tackle various aspects of the complex issues inherent in these technologies.

At the beginning of the research project, particular attention was given to the process variables and the quality of the final products. In this initial study, we examined the relationship between process variables and quality, specifically focusing on the identification of key variables that significantly influence the quality outcomes. In this particular scenario, our focus was on employing a variable selection model that could effectively take into consideration the high correlation present in the data. To address this challenge, a method has been developed to identify relevant variables for quality assessment while accounting for their strong correlation. The method has been compared with numerous variable selection algorithms, showing excellent results and outperforming all of its competitors. By employing this methodology in the manufacturing case, we successfully investigated the relationship between process data and quality of the products.

In the context of process optimisation and the pursuit of obtaining a high-quality level of the final products, it is necessary to establish a causal relationship between the input parameters and output data. Controlled experiments are commonly employed in order to establish causal relationships. In the context of additive manufacturing, emerging technologies exhibit increased complexity that introduces an abundance of input parameters, rendering experimentation an unfeasible opportunity. In this context, we have thus implemented an approach that aims to provide valuable insights

into the initial experimentation stage. Specifically, this strategy involves the limitation of the input space and a selection of a small number of parameters. This approach applied in the practical case has allowed us to establish two input variables that influence the mechanical quality of the final products.

A significant amount of the available data regarding 3D printers pertains to process-related information. Process data refers to the data that is obtained throughout the production process for every individual micro-layer that is deposited on the building bulk. The relationship between final products and the observations obtained during their manufacture creates a grouping structure within the mode related to the observations. An approach for unsupervised decomposition of multi-group three-way arrays has been developed. This methodology aims to build a model and extract the model loadings that are explanatory of the common structure shared by the groups. The loadings in this context serve as explanatory factors, revealing the underlying common structure among the different groups in the data set. This method has been applied in the context of additive manufacturing, providing valuable insights for enhancing understanding of the process.

A natural extension of the previous model is transitioning from an unsupervised model to a supervised model by taking into account the quality variables. The model is an extension of the Partial Least Squares (PLS) model to the case of multi-group three-way arrays. This model provides various benefits due to its ability to effectively exploit the intricate information included in the multi-group three-way arrays and the correlation with the response variable. The model is also a predictive model, capable of not only understanding the complex interconnections between input and output data but also forecasting the outcomes of these linkages. In the context of new technologies, characterised by complex variables and diverse dynamics, this particular model represents considerable potential. The proposed model is in its early stages but has significant potential to make a valuable contribution to technological progress.

A key component of the project is working in collaboration with the other components of the broader initiative, namely, three doctoral candidates engaged in diverse activities related to the 3D printers. Together we engaged in several collaborative projects that required the collective contribution of each team member's diverse expertise, different domain knowledge and prior experience within the realm of research. In Chapter 7, some of these collaborative projects were presented. The first collaboration project regards a laser profiling system which opens the possibility for online quality monitoring. The second collaborative project regards the investigation of dimensional faults encountered in the final products. These initiatives are currently ongoing projects and make valuable contributions to the overall enhancement and improvement of the processes.

The new 3D printing technologies have inspired and driven the development of new statistical and machine learning models for data analysis. The successful implementation of the new methodologies developed in the course of this doctoral research enabled us to effectively address a multitude of challenges and objectives. The established methodologies are expected to enhance the spectrum of data analysis models and provide support to academics and practitioners in addressing their analytical

challenges. The developed models have been utilised in the context of the additive manufacturing field. The findings have been examined in collaboration with process engineers, contributing to the advancement of the knowledge and understanding of the production processes. Nevertheless, it is important to acknowledge that the journey of process optimization is ongoing since there is a significant demand for more advancements and improvements in enhancing the reproducibility of the machinery and overall optimisation and stability.

8.1 Future work

The close collaboration with other PhD students of the group fostered the opportunity to engage in many projects utilising diverse methodologies and interfacing with models from many disciplines. Throughout the course of these projects, a significant emphasis has been placed on comparative analysis. Numerous discussions centred around the insights gained from each model, shedding light on their respective underlying principles and conceptual foundations. The results obtained from different models, each representing a unique perspective, have been systematically evaluated and contrasted. This comparative assessment has led to a deeper understanding of each model type's strengths, limitations, and potential applications. These interactions have contributed to the improvement of individual comprehension and also stimulated the amalgamation of novel concepts and methodologies by fostering the interchange of varied perspectives.

The dynamic of the collaborative group has facilitated the use of several techniques and methodologies, and numerous models, both data-driven and physics-based, have been developed. The close collaboration between the data and physically driven findings has enabled the potential integration of not only the final outcomes but also the underlying models. This paves the way for the development of new models that can be a combination or hybrid models between both worlds. The essence of collaboration between these distinct projects gives the opportunity for the development of a remarkable synergy. This collaborative platform provides an interesting environment for the development and implementation of innovative models that effectively integrate components from both areas.

Data-driven models include many models ranging from statistical, machine learning to artificial intelligence models. Statistical models aim to characterize and capture the underlying patterns and distributions within the collected data. Statistical modelling allows for hypothesis testing, parameter estimation, and prediction, facilitating a quantitative understanding of the process dynamics. Machine learning techniques have emerged as powerful tools due to their ability to handle complex and non-linear relationships. Machine learning and AI models such as artificial neural networks, decision trees, support vector machines and ensemble methods aim to uncover hidden patterns and make accurate predictions based on the available data. These models excel at capturing intricate dependencies and non-linear dynamics, making them

suitable for diverse applications within process data analysis.

On the other hand, physics-based models or as often referred to as first-principles models or mechanistic models, incorporate domain knowledge and fundamental principles of the underlying physical processes into the modelling framework. These models are often derived from first principles, mathematical equations, or physical laws that govern the system being studied. By combining process knowledge with experimental data, physics-based models can simulate and predict the behaviour of the production process accurately. They provide valuable insights into the physical phenomena and mechanisms that drive the process, enabling deeper understanding and facilitating process optimization.

The integration of physics-driven models with data-driven models establishes a foundation for the development of hybrid models that incorporate the advantageous qualities and collaborative effects inherent in each individual technique. Integrating physics-based models, which are based on fundamental principles and supported by theoretical frameworks, with statistical models, which extract patterns and correlations from real-world data, leads to a comprehensive explanation of complex events. This type of model has the potential to exploit the precision of theoretical framework while using the predictive power obtained from empirical data. These hybrid models could, by design, synthesize insights and principles from data-driven and physics-based approaches, resulting in an approach that incorporates the strengths and overcomes the limits of both methodologies.

This journey towards hybrid models, although ongoing, holds immense promise. The resultant models have the potential to provide a complete and all-encompassing representation of the complex systems under investigation. By integrating empirical data with established physical laws, these models are expected to provide valuable insights that would be difficult to get using either technique independently.

Bibliography

[1] D. V. Rosato and M. G. Rosato, *Injection molding handbook*. Springer Science & Business Media, 2012.

[2] Z. Chen and L.-S. Turng, "A review of current developments in process and quality control for injection molding," *Advances in Polymer Technology: Journal of the Polymer Processing Institute*, vol. 24, no. 3, pp. 165–182, 2005.

[3] A. Patil, A. Patel, and R. Purohit, "An overview of polymeric materials for automotive applications," *Materials Today: Proceedings*, vol. 4, no. 2, pp. 3807–3815, 2017.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] M. Segal and Y. Xiao, "Multivariate random forests," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 1, no. 1, pp. 80–87, 2011.

[6] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The annals of applied statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[7] M. Zhu, J. Xia, X. Jin, M. Yan, G. Cai, J. Yan, and G. Ning, "Class weights random forest algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018.

[8] L. Auret and C. Aldrich, "Interpretation of nonlinear relationships between process variables by use of random forests," *Minerals Engineering*, vol. 35, pp. 27–42, 2012.

[9] L. Guo, Y. Ma, B. Cukic, and H. Singh, "Robust prediction of fault-proneness by random forests," in *15th international symposium on software reliability engineering*, pp. 417–428, IEEE, 2004.

[10] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–13, 2010.

[11] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, 2017.

[12] R. Nilsson, J. M. Pena, J. Björkegren, and J. Tegnér, "Consistent feature selection for pattern recognition in polynomial time," *The Journal of Machine Learning Research*, vol. 8, pp. 589–612, 2007.

[13] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1200–1205, Ieee, 2015.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[16] V. Kumar and S. Minz, "Feature selection: a literature review," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[17] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, vol. 43, no. 5, pp. 2055 – 2085, 2015.

[18] E. J. Candès, Y. Fan, L. Janson, and J. Lv, *Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection*, vol. 1610. Department of Statistics, Stanford University Stanford, CA, USA, 2016.

[19] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[20] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Vsurf: an r package for variable selection using random forests," *The R Journal*, vol. 7, no. 2, pp. 19–33, 2015.

[21] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[22] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," 2015.

[23] E. J. Candes, Y. Fan, L. Janson, and J. Lv, "Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection; department of statistics," 2016.

[24] "Variable selection with knockoffs description." `https://web.stanford.edu/group/candes/knockoffs/index.html`.

[25] L. R. Tucker *et al.*, "The extension of factor analysis to three-dimensional matrices," *Contributions to mathematical psychology*, vol. 110119, 1964.

[26] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[27] R. Bro, "Parafac. tutorial and applications," *Chemometrics and intelligent laboratory systems*, vol. 38, no. 2, pp. 149–171, 1997.

[28] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 24–40, 2011.

[29] R. A. Harshman *et al.*, "Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis," 1970.

[30] J.-h. Jiang, H.-l. Wu, Y. Li, and R.-q. Yu, "Three-way data resolution by alternating slice-wise diagonalization (asd) method," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 14, no. 1, pp. 15–36, 2000.

[31] P. Paatero, "A weighted non-negative least squares algorithm for three-way 'parafac'factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 223–242, 1997.

[32] N. Faber, L. Buydens, and G. Kateman, "Generalized rank annihilation method. i: Derivation of eigenvalue problems," *Journal of chemometrics*, vol. 8, no. 2, pp. 147–154, 1994.

$$\Theta^{\sqrt{17}} + \Omega \int \delta e^{i\pi} = -1$$

$$\{2.7182818284\} \mathring{\lambda}$$

ϑφερτυθιοπσδφγηξκλ

$$\chi^2 \sum_{!} \gg \approx$$