Knowledge extraction from SME data for the implementation of PHM process

Nabil Omri

SPIM

Thèse de Doctorat

UBFC

école doctorale SPIM sciences pour l'ingénieur et microtechniques

UNIVERSITE DE BOURGOGNE FRANCHE - COMTE

Thèse préparée à l'Université de Bourgogne Franche-Comté et présentée par

M. Nabil OMRI

pour obtenir le

Grade de Docteur de l'Université de Bourgogne Franche-Comté

Spécialité : **Sciences pour l'ingénieur**

Extraction de connaissances à partir des données d'une PME en vue de l'implémentation du PHM

Knowledge extraction from SME data for the implementation of PHM process

Unité de Recherche : FEMTO-ST, UMR CNRS 6174

Soutenue le 22 avril 2021 devant le Jury :

M. Vairac Pascal	Professeur des Universités, ENSMM Besançon	Président
M. Lamouri Samir	Professeur des Universités, CNAM Paris	Rapporteur
M. Pérès François	Professeur des Universités, INP ENIT Toulouse	Rapporteur
Mme Ribot Pauline	Maître de conférences, LAAS-CNRS	Examinatrice
Mme Hajri Sonia	Professeur des Universités, INSAT Tunis	Examinatrice
M. Zerhouni Noureddine	Professeur des Universités, ENSMM Besançon	Directeur de thèse
Mme Al Masry Zeina	Maître de conférences, ENSMM Besançon	Codirectrice de thèse
M. Giampiccolo Sylvian	Directeur général SCODER, Pirey	Invité
M. Mairot Nicolas	Directeur SCM SCODER, Pirey	Invité

For my parents.
For Imen and my little I.
For my brothers and my sisters.
For Rawssen, Rassil, Roumaysaa and Aws.
For Brigitte and Farhat.

Acknowledgments

This thesis has been a rich experience for me that cannot be completed without thanking the people who have supervised, helped and supported me during the last three years.

I would like to thank first of all my supervisors, Prof. Noureddine ZERHOUNI, Dr. Zeina AL MASRY from the FEMTO-st institute and Mr. Sylvian GIAMPICCOLO and Mr. Nicolas MAIROT from the SCODER company, for the confidence they have given me, their constant support and follow-up, their precious advice and their availability throughout this thesis.

I warmly thank the members of my jury, the rapporteurs Prof. Samir LAMOURI and Prof. François PERES, as well as the examiners Prof. Pascal VAIRAC, Prof. Sonia HAJRI and Dr. Pauline RIBOT, for having done me the honor of reporting and examining my work and for having traveled to attend my defense.

I would like to thank all the staff of the SCODER company for the pleasant working atmosphere and the daily exchanges. A big thank you to Audrey, Béatrice, Romuald, Pascale, Cindey, Hervy, Gille, Bernard, Denis, Jean-Pierre, Cédric, Michel, Xavier, Moustapha Christophe, David and Patrick for offering me their friendship and for all that we lived together.

I thank all the staff of the AS2M department of the FEMTO-ST institute for the pleasant working atmosphere and the daily exchanges. A big thank you to Amin, Omar, Safa, Khaled, Vladémir and Chifaa for offering me their friendship and for all that we have lived together.

Finally, I would like to thank all the people who participated directly or indirectly in the success of this work.

Nabil Omri

Contents

List of Figures

List of Tables

Publications

All the approaches and results presented in this thesis have been published in international journals and conferences:

International journals

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., Zerhouni, N. (2021). Towards an adapted PHM approach: Data quality requirements methodology for fault detection applications. *Computers in Industry, 127, 103414.*

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni,N. (2020). Industrial data management strategy towards an SME-oriented PHM. *Journal of Manufacturing Systems, 56:23–36.*

International conferences

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni,N. (2020). Data quality requirements methodology for an adapted PHM implementation. *In the International Conference on Communication and Intelligent Systems (India-2020).*

Omri, N., Al Masry, Z., Giampiccolo, S., Mairot, N., and Zerhouni,N. (2019). Data management requirements for PHM implementation in SMEs. *In 2019 Prognostics and System Health Management Conference (PHM-Paris), pages 232–238.IEEE.*

National conferences

Omri, N., Al Masry, Z., and Zerhouni,N. (2019). A Data-Driven PHM approach for SMEs. *In 16ème Colloque National S-mart.*

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni,N. (2020). L'Industrie 4.0 à la portée des PME. *Published in the MICADO newspaper N°5 2020.*

Journal and conference articles submitted or in preparation.

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni,N. (2021). Une approche pour la formalisation des connaissances en PHM: application aux PME. *Submitted to the CIGI Qualita21 conference.*

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni,N. (2021). X-PHM: A user friendly PHM framework towards an SME-adapted PHM. *Submitted to the 54th CMS CIRP 2021 proceedings.*

Acronymes & Notations

Acronymes

PHM	*Prognostics and Health Management*
SME	*Small and meduim-sized Enterprise*
MNE	*Multi-National Enterprise*
DQ	*Data Quality*
DQD	*Data Quality Dimension*
DS2	*Scoder Data System*
ROI	*Return On Investment*
CBR	*Case-Based Reasoning*
ANN	*Artificial Neural Network*

Notations

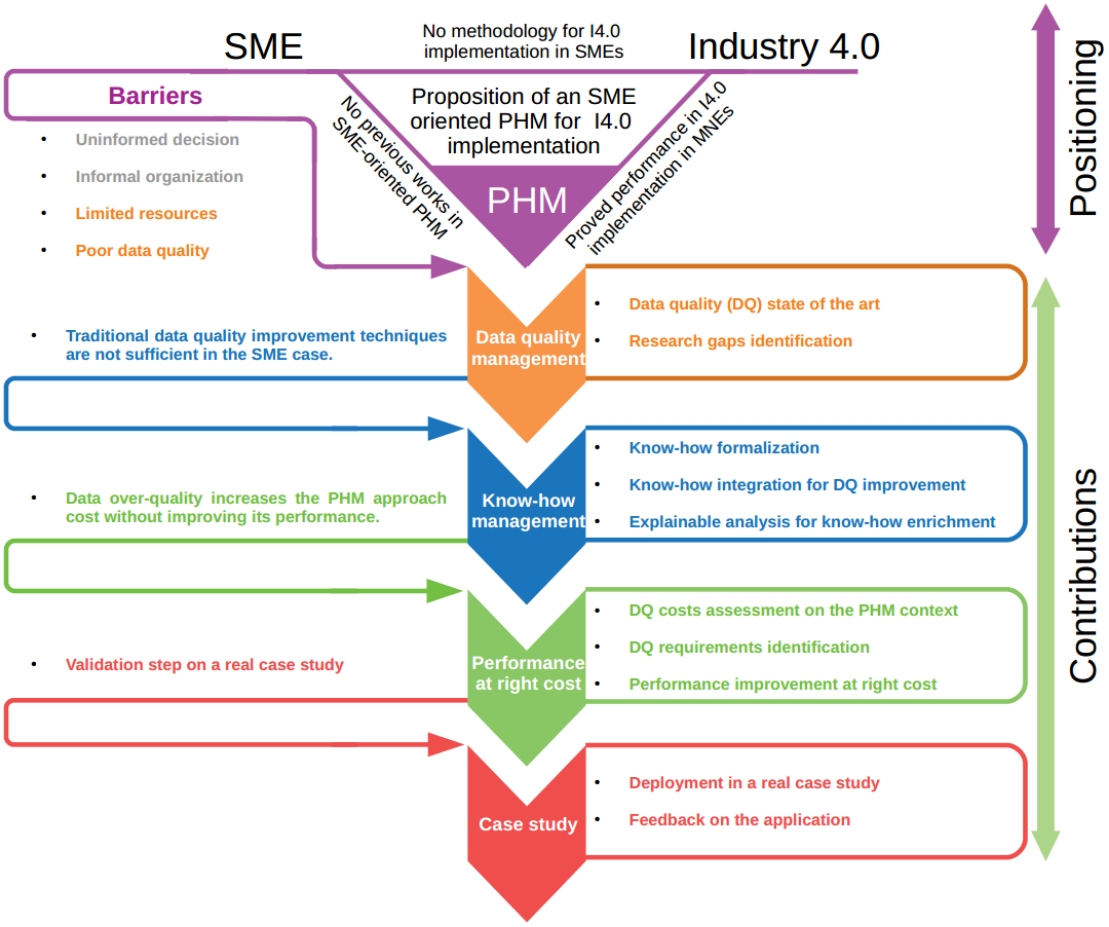
Symbol	Description
$\sum$:	The studied system
Det:	Detectability state of the system $\sum$
O:	Observability state of the system $\sum$
Q:	Data quality of the dataset
P:	Performance of the used detectability algorithm
GQ:	Global data quality
LQ:	Local data quality
X_i:	Features that describe the system $\sum$ for $i = 1, \ldots, n$
Q_i:	Data quality of a features X_i
q_{Im}:	Imbalanced data ratio

continued on next page

continued from previous page	
Symbol	Description
q_{i1}:	Missing data ratio for a feature X_i
q_{i2}:	Noisy data ratio for a feature X_i
w_i:	Importance weight of the feature X_i
CD:	Cost of a negative detection
CI:	Needed cost to assess an imbalance ratio level
CM_i:	Required cost to assess a missing data ratio level
CN_i:	Required cost to assess a noisy data ratio level
$\mid . \mid$:	Cardinality of the data space

TABLE 1: Notations summary

Graphical abstract



Introduction

"A problem is a chance for you to do your best."

-Duke Ellington

The industrial domain has evolved rapidly over time through various industrial revolutions, from the steam engine to digitization technology. Throughout these changes, small and medium-sized enterprises (SMEs) have always played a key role in the global economy as well as in the French economy. Indeed, they represent 90% of french companies with employability of 6.3 million employees, and 43% of added value [France-industrie, 2020]. Despite these impressive statistics, french SMEs, like most SMEs worldwide, face several difficulties on the road to their digital transformation [Omri et al., 2020]. The digital transformation in SMEs is limited by the unavailability of resources and the particular SMEs organization [Omri et al., 2019]. Compared to multinational enterprises (MNEs), SMEs present a lack in terms of the **available resources**. MNEs have the privilege of possessing all the required resources (human, technological and financial resources) for a development project. Thus, MNEs are better able to lead the research and development domain than SMEs. This lack reflects the growing gap between small and large groups. The consequence of this gap is the low competitiveness of SMEs [Omri, 2020]. In this context, the digitization process and the integration of Industry 4.0 technologies seem essential for the development and growth of SMEs [Moeuf, 2018, Omri et al., 2020]. However, many obstacles limit their transfer to Industry 4.0 [Mittal et al., 2018].

This thesis work is part of a collaboration between the SCODER company and the FEMTO-ST institute. SCODER is a french SME installed near Besançon and specialized in ultra-precise stamping for automotive applications. The addressed problem proposes upstream research, innovative, brings together several topics of the institute. It has been tackled head-on in all disciplines by proposing an original approach driven by industrial needs. Concretely, the subject is about developing an approach facilitating the integration of Industry 4.0 technologies within SMEs. This project deals with many

disciplines such as data acquisition, data analysis, and cost optimization. These disciplines are widely studied in the literature. However, it remains a crucial research gap in this context: data quality management in digitization projects. For that, our interest is specifically focused on data quality management in SME cases.

The starting point of the proposed approach is based on a preliminary bibliography study reinforced by field observations. This study allowed to define the required parameters for the Industry 4.0 integration within SMEs. These parameters (see Figure 1) can be summarized in three points: data structuring, knowledge management and performance improvement at the right cost.

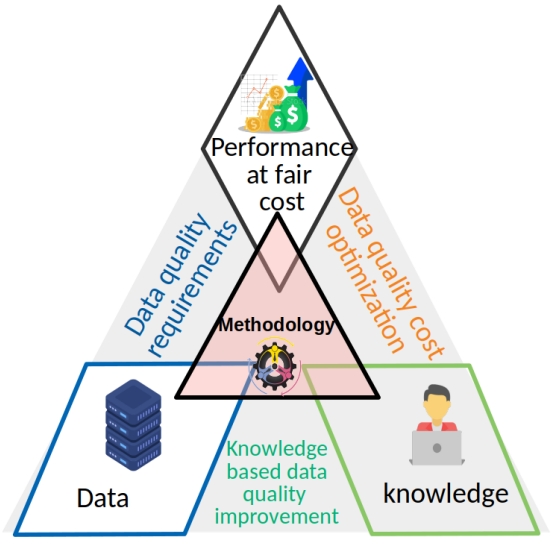


Figure 1 - Required parameters for the integration of Industry 4.0 within SMEs.

- **Data:** SMEs have a large amount of data in the form of monitoring logs and know-how. However, only a small amount of this data is digitally recorded. And even this tiny quantity presents several quality problems. Therefore, the real challenge is to valorize these data and extract knowledge from them while taking into account these quality problems.
- **Knowledge:** The advantage in SMEs is that operators are close to the process, which allows them to identify their needs and prioritize them in terms of urgency. The staff's versatility allows them to analyze each problem regarding its impact on the entire production process. Thus, the valorization of this knowledge is an essential element for successfully integrate Industry 4.0 within SMEs.
- **Performance at right cost:** The ultimate goal of the Industry 4.0 concept is to improve industrial performance and maximize profits. However, the economic justification for these improvements is little discussed. Thus, it is essential to propose

a methodology for enhancing industrial performance at the right cost to promote the Industry 4.0 integration within SMEs.

In this context, the prognostics and health management (PHM) concept seems adequate for Industry 4.0 integration within SMEs [Omri et al., 2020]. PHM is an emerging concept in the industrial domain that brings together several disciplines for performance improvement. The interest of the PHM lies in the fact that it is a concept that starts from data through analysis and ends with a decision for performance improvement. Thus, we propose in this thesis to adopt the PHM concept as a solution to facilitate Industry 4.0 integration within SMEs. However, the PHM concept is not widely applied in the SMEs domain [Omri et al., 2019]. Thus, additional efforts are required to adopt this concept in SME cases, especially concerning the available data quality. Therefore, a large part of this work has been devoted to developing these points and proposing adequate solutions to facilitate digital transfer in SMEs through an adapted data-oriented PHM approach. All the work carried out in this thesis is validated and encapsulated in Scoder Data System (DS2) software. DS2 is an application developed during this thesis and aims to facilitate the Industry 4.0 integration within the SCODER company. Thus, a part of this manuscript is devoted to the presentation of this software.

This thesis is articulated in five chapters:

Chapter 1 underlines the SCODER industrial problem statement in the Industry 4.0 context. The industrial problem statement emerges from a concrete use case, materialized by a stamping line in the SCODER factory. It addresses the challenge of implementing industry 4.0 technologies within SMEs. Thus, this chapter's main contributions consist of identifying the barriers that limit the Industry 4.0 implementation within SMEs. Based on these barriers, an adapted PHM approach is proposed to implement Industry 4.0 within SMEs. This approach is applied in the SCODER company, and the research gaps were identified. One of the most critical gaps in this domain is the data quality management issue, which consists of this thesis's backbone.

Chapter 2 consists of state of the art on data quality management. Thus, an overview of data quality definitions, dimensions, and metrics is presented. Moreover, the most encountered data quality problems in the PHM context are studied and presented as the proposed approach's focal point. To improve data quality, an overview of the adopted improvement techniques is proposed. Based on this literature review, we identify the data quality improvement within SMEs as the central research gap, which we suggest filling in the next chapter.

Chapter 3 deals with the data quality improvement issues in the SMEs context. One possible solution consists of formalizing the operators' know-how and its utilization for data quality improvement. For this purpose, human knowledge types are defined, and

their integration modes in the data management process are discussed. Based on this study, a closed-loop (data quality - knowledge) is developed for data quality improvement and human knowledge enrichment. The proposed approach is applied in the SCODER case study to validate it and assess its applicability.

Chapter 4 deals with the data quality cost optimization. To do, the data quality impact on the fault detection task of the PHM process is formalized. This formalization is used to propose an empiric metric to quantify this impact. Based on this empirical metric, different scenarios for data quality cost optimization are proposed. One of these scenarios is applied to the SCODER company to improve the industrial performance at the right cost.

Chapter 5 consists a global validation step of this thesis work. The developed approach is applied in the SCODER company. To facilitate the implementation, all the steps of this methodology have been encapsulated in SCODER Data System (DS2) software. Thus, this chapter has been devoted to present this software and its functionalities through a real case study of metal coils assignment optimization. The installation steps are described, and the obtained results are reported.

Finally, this manuscript ends with a conclusion of this thesis work and discusses the research lines considered as perspectives to this work.

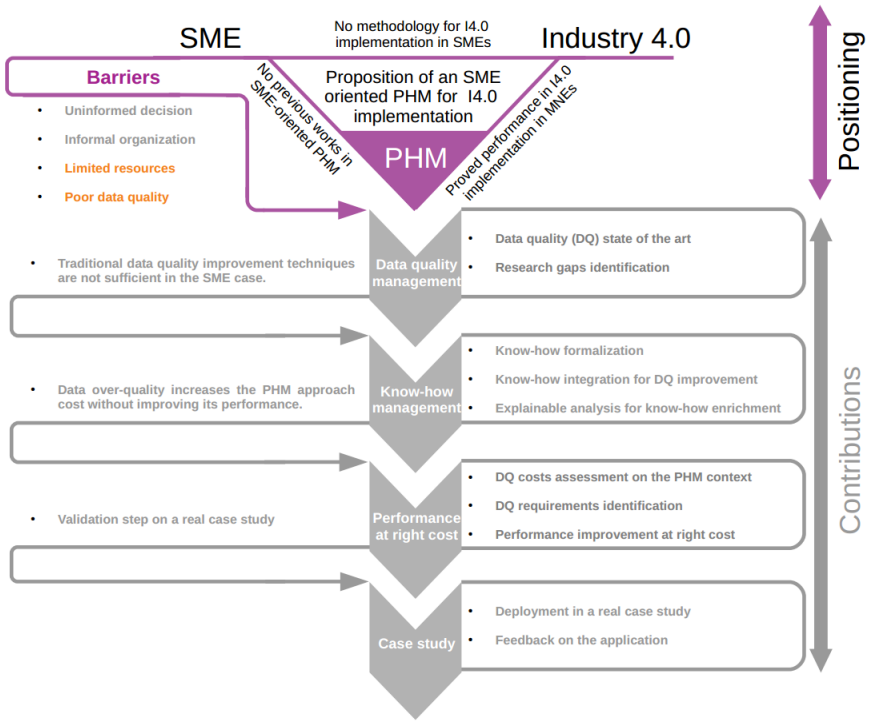
Chapter 1

Data driven PHM: a key to implement Industry 4.0 in SME

"A big business starts small."

-Richard Branson

Graphical abstract.



Contents

Contributions

This chapter underlines the SCODER industrial problem statement, which emerges from a concrete use case, materialized by a stamping line in the SCODER factory. It addresses the challenge of implementing industry 4.0 technologies within SMEs. Thus, this chapter's main contributions consist of identifying the barriers that limit the Industry 4.0 implementation within SMEs. Based on these barriers, an adapted prognostics and health management (PHM) approach is proposed to implement Industry 4.0 within SMEs. This novel approach is applied in the SCODER company, and the research gaps were identified. One of the most critical gaps in this domain is the data quality management issue, which consists of this thesis's backbone.

1.1 Introduction

The fourth industrial revolution is derived from advances in digitization and data analysis disciplines to make plants smarter and more efficient. However, an adapted approach for Industry 4.0 implementation in small and medium-sized enterprises (SMEs) has not been yet discussed. This research gap is due to SMEs' specificities and the lack of a clear methodology that respects these specificities. This chapter proposes general definitions of SMEs and the Industry 4.0 concept while focusing on their characteristics and inadequacies. To deal with these inadequacies, we propose to adapt the prognostics and health management (PHM) process and use it to implement Industry 4.0 in SMEs and thus facilitate their digital transformation. The proposed SME-oriented PHM approach is applied in the SCODER company, which is a French SME. The conducted literature review combined with this real-world application allows identifying the research gaps that need to be filled to implement Industry 4.0 within SMEs successfully.

The remainder of this chapter is organized into four sections. Section 1.2 presents this thesis's scientific positioning regarding the Industry 4.0 implementation within SMEs. Section 1.3 introduces the SMEs characteristics. Section 1.4 presents the Industry 4.0 concept while discussing the barriers that limit its integration within SMEs. Sections 1.5 and 1.6 introduce the PHM concept as a key-enabler to implement Industry 4.0 in SMEs. The first results of applying this approach in the SCODER company are detailed in Section 1.7 while presenting the different research gaps that need to be filled to successfully implement Industry 4.0 within SMEs using an SME-oriented PHM process. Finally, Section 1.8 concludes this chapter.

1.2 From the SCODER case to a general SMEs problem

This section highlights the SCODER industrial issue within the context of a digitization project led by the SCODER company to integrate Industry 4.0. The industrial problem

statement emerges from a concrete use case, materialized by many cutting and stamping lines. It addresses the challenge of integrating industry 4.0 technologies within an SME with limited resources (financial resources in particular). Thus, the thesis deals with two major challenges: the proposition of a methodology to integrate Industry 4.0 within SMEs and the consideration of the limited resources within SMEs, requiring the optimization of installation costs.

1.2.1 Presentation of the SCODER company

SCODER is founded in 1954 by toolmakers from Besançon. Then, in 1988, the company joined the R. BOURGEOIS group, world number 3 in cutting engine plates [SCODER, 2021]. In 1998, SCODER moved to its current premises in Pirey (see Figure 1.1). In 2006, the company created a subsidiary in the Czech Republic for the assembly of primary parts cut in Pirey. The company now has around 120 employees spread over three production teams in Pirey, and about 30 in the Czech Republic [SCODER, 2021].

FIGURE 1.1: Presentation of the SCODER company.

The company has a rich experience of more than 60 years in cutting and certifications to ISO 9001 and ISO TS 16949 standards. In addition, it has a fleet of machines made up of presses ranging from 75T to 800T. This allows the company to offer many intricate parts to the automotive industry. This sector represents 95% of the company's production. Thus, SCODER produces different types of cut pieces for various uses. SCODER is also specialized in the design and production of cutting tools. This asset allows the company to control the entire process. Initially, the company carries out a co-design with the customer to optimize the cost and the quality of the product. The tools are designed, manufactured, and developed in the company's workshops. Besides, the company performs maintenance on its cutting tools. The company also assembles some of the parts it cuts. It is carried out by welding robots of different technologies (MAG, PLASMA, TOPIG). The latter tends to become universal both in terms of quality and safety [SCODER, 2021].

1.2.2 Presentation of the studied system

The thesis's research object is a sheet metal cutting line, shown in Figure 1.2, of which more details on the press used are shown in Figure 1.3.

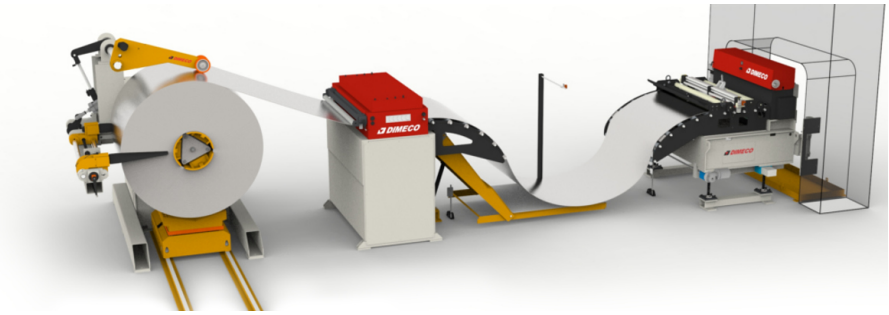


FIGURE 1.2: Presentation of a cutting line.

This cutting line is installed in the SCODER plant in Pirey, and it is used for high-precision cutting of parts intended for the automotive sector. Such production line is composed of three large blocks: *(i) Mandrel, (ii) Rectifier, and (iii) Press.*

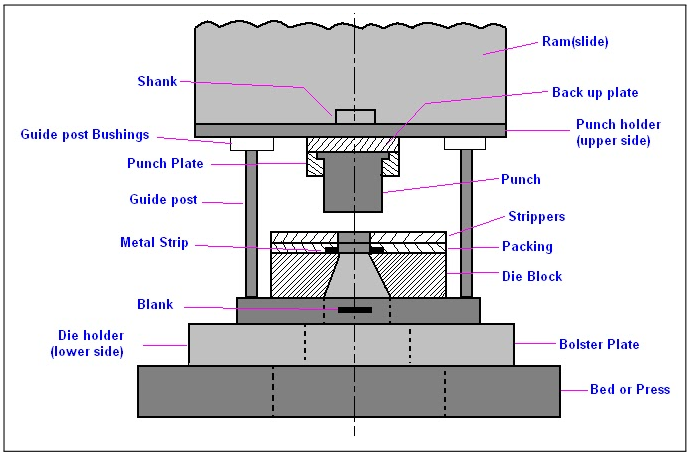


FIGURE 1.3: Presentation of a die-cutting press.

The centerpiece of this production line is the press. Indeed, it is responsible for the cutting operation. A press is a machine used for the deformation of sheet metal. Figure 1.3 shows the different components of a mechanical press. Usually, a press is composed of a base, a frame, guide columns, motors for the generation of force, and systems for the management and control of the deformation phase. The fixed part of the die is clamped on the base, while the movable element is connected to the press housing base. The frame is used to counterbalance the force imposed on the mold, while the columns or guides ensure perfect alignment of the die parts.

1.2.3 Problem and objectives

In this paragraph, we present the problem of this study. Indeed, a cutting line is exposed to several production anomalies that result in production shutdowns or quality problems. To ensure stable production in terms of productivity and part quality, one track is to install Industry 4.0 technologies within these cutting lines. However, an adapted approach for Industry 4.0 implementation in small and medium-sized enterprises (SMEs) has not been yet discussed. This research gap is due to SMEs' specificities and the lack of a clear methodology that respects these specificities. Thus, the SCODER company's motivation to install Industry 4.0 technologies as part of its digital transformation leads to express a research problem raised by the daily concerns of the workshops. This research problem concerns:

- The identification of the SMEs' characteristics and their limits regarding the integration of Industry 4.0.
- The proposition of an adequate approach to integrate Industry 4.0 within SMEs.

The remainder of this chapter is dedicated to studying the SME characteristics and the barriers that limit their integration of Industry 4.0 technologies.

1.3 Small and medium-sized enterprises

The industrial domain has evolved rapidly over time through the various industrial revolutions, from steam engines to cloud computing technology [Wang et al., 2018]. Throughout these changes, SMEs have always played an essential role in global business. However, there is no standard definition of SMEs.

1.3.1 SME definition

SME's definition varies from country to country. In China, SMEs are defined, according to the Law of the People's Republic of China on the Promotion of Small and Medium-sized Enterprises [China.org.cn, 2021], as companies that "have a relatively small size in personnel and scope of business." The standards for classifying small and medium enterprises are formulated by the relevant departments of the State Council. Identifying a company as a micro, small or medium-sized enterprise depends on a series of variables such as the industry it belongs to, its operating income, its total assets, and its number of employees [Guo et al., 2019]. In Canada, SMEs are defined as all firms that employ between 1 and 499 people and whose turnover does not exceed C$ 50 million [Canada, 2021].

In this work, we adopt the French definition of SMEs, which is governed by the Law on the Modernization of the Economy of August 4, 2008 (Application Decree No. 2008-1354, Article 51) [France-industrie, 2020]. SMEs include companies with more than ten

employees and fewer than 250 employees and with annual turnover less than 50 million euros or a total balance sheet not exceeding 43 million euros.

According to the 2019 SBA Fact Sheet [SBA2019, 2021], the French manufacturing sector generated weak SME added value growth of only 0.7% in 2014-2018. SME employment declined by 1.3% in the same period, and the fall was even steeper in large firms, at 7.1%. Conversely, the added value of large firms rose by 13.6%. In 2017-2018, SME added value dropped by 1.0%, while employment increased by 1.4%. This sluggish overall manufacturing performance can be explained by a decline in production, most notably in the manufacturing of durable consumer goods.

In 2012, Louis GALLOIS submitted his report to the French Prime Minister on the competitiveness of French industry [Louis, 2012]. He noted that SMEs suffered from significant weaknesses, including lack of equity capital, difficulty opening up wealth, fear of management, and fear of investment risk. These weaknesses are accentuated by the industry's performance over the last ten years. Between 2007 and 2012, the number of SME failures has increased from 3,100 per year to 4,600 per year. Over the same period, SMEs saw their gross margin shrink by more than six percentage points, from more than 25% in 2007 to less than 19% in 2009.

However, SMEs should not merely be considered by the statistics relating to their definition. They represent organizations with specific characteristics that we describe in the following subsections.

1.3.2 SME characteristics

SMEs are a special type of companies that are widely studied in the literature [Moeuf et al., 2018, Omri et al., 2019, Mittal et al., 2018, Omri et al., 2019]. In this paragraph, we propose analyzing these studies to identify these companies and their particular environment by differentiating them from big corporations. Based on this analysis, the different opportunities and challenges faced by SMEs are identified and discussed. In [Mittal et al., 2018], the authors propose eight clusters to characterize SMEs. These clusters are finance, technical resource availability, product specialization, standards, organizational culture, employee participation, alliances, and collaboration. In the same context, Moeuf et al. [Moeuf et al., 2018] propose to study the SMEs managerial characteristics. Thus, the authors propose to situate the SME within companies' growth process and detail the coordination modes that govern these companies. In [Omri et al., 2019], the authors propose to aggregate these clusters into resource-based, organization-based and methodological characteristics.

In this thesis, we propose to adopt the clustering proposed by [Omri et al., 2019]. For that, SMEs characteristics are discussed according to three clusters: resource-based, organization-based and methodological characteristics (See Figure 1.4).

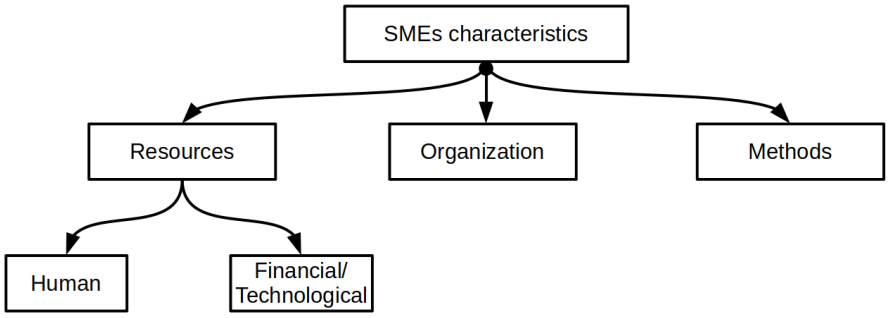


FIGURE 1.4: The proposed SMEs characteristics clustering.

Resource-based characteristics: Resource availability is one of the essential features that can be used to describe a company. In this context, Subrahmanya [Subrahmanya, 2015] classifies these resources into four main categories (physical, human, technological, and reputational). We here focus on human and technical resources. One should note that technological resources are highly dependent on financial resources. For that, financial resources are described as a part of the technological entity.

- **Human resources:** In [Subrahmanya, 2015], authors affirm that unskilled workers join SMEs for a short period, and when they are fully equipped with the "required skills", they leave to join large enterprises. Consequently, advanced tasks cannot be supported by SMEs. Moreover, the received opportunities by SME employees are very different from those in multinational enterprises (MNEs) [Mittal et al., 2018]. SME employees are known to be "jack-of-all-trades" [Bublitz and Noseleit, 2014], which prevents them from developing high skills in a particular domain [Dombrowski et al., 2010]. However, this polyvalence can result in a better understanding of the overall production process and its interactions.

- **Financial and technological resources:** Financial resource availability is an essential factor for the development of any company. In this context, SMEs are reported to be financially constrained [Mittal et al., 2018]. They are often owned by families, which implies several weaknesses related to small businesses, such as the lack of guarantees and investors' absence. These risks are at the root of SMEs' capital constraints compared to MNEs [Jasra et al., 2011].

  This lack of funding leads to difficulties in adopting new technologies and the upgrading of existing ones [Dangayach and Deshmukh, 2005]. In this context, only 5 % of equipment in companies are digitally controlled [Waurzyniak, 2015]. This problem is mainly related to the lack of adequate measurement technologies to meet the manufacturing environment's needs. In this context, SMEs can be considered as underserved markets [Helu et al., 2015].

Organization-based characteristics: In this paragraph, we highlight the organizational characteristics of SMEs. The firm's organization is less formal, and communication

between management and staff is close and informal [Durst and Bruns, 2018]. Also, organizational culture is generally insufficiently flexible in experimenting and considering new initiatives [Van de Vrande et al., 2009]. From the documentation point of view, the information about the production process is limited in operators' know-how without any documentation. In a case study of two SMEs, Boden et al. [Boden et al., 2012] reported that only very little documentation is available in these companies.

Moreover, it never goes into details on how things are implemented. The information process is not digitized, but it is based on documents edited by operators [Omri et al., 2019]. For that purpose, the corporate memory (if it exists) is created manually by workers. Besides, SME owners are accused of being "strategically myopic" and lacking the "long-term vision" [Wang et al., 2007]. As a result, medium and long-term strategies are rare in SMEs. Moreover, SME's decisions are not as informed and based on the owner's feeling, which involves high levels of uncertainty [Salles, 2006].

When it comes to collaborations and alliances, SMEs lack alliances with universities and other research institutions, which prevents them from knowing these institutions [Mittal et al., 2018]. Moreover, SMEs manage a limited list of product which reduce the number of suppliers, and consequently, they depend strongly on them [Singh et al., 2007].

Methodological characteristics: Due to the limited technical and financial resources, SMEs' research and development domains are not very advanced. Still, their hard work leads to highly specialized products that can differentiate SMEs from their competitors [Julien and Ramangalahy, 2003b]. The lack of awareness and resources compared to MNEs makes SMEs' survival difficult [Lee et al., 2010]. MNEs strictly obey ISO standards; however, the presence of these standards in SMEs is rare. It is partly due to the resources required to prepare and pass the certifications [Brown et al., 1998]. Therefore, SMEs need to consider the industrial standards. Study [Blind and Mangelsdorf, 2012] conducted in electrical engineering and machinery micro firms in Germany showed that SMEs are interested in accessing the knowledge gained by MNEs. Still, they think that standardization may disclose their essential information to the competitors.

1.3.3 Synthesis on SMEs characteristics compared to MNEs

Based on the previously discussed attributes, this paragraph proposes a synthesis of SMEs' characteristics compared to MNEs. Table 1.1 summarizes the main differences between big companies and SMEs according to the previously defined clusters (Resource, Organization, and Method).

Compared to MNEs, SMEs present a lack in terms of the **available resources**. MNEs have the privilege of possessing all the required resources (human, technological and financial resources) for a development project. Moreover, international relations and inter-organizational cooperation are more developed within MNEs [Omri et al., 2019]. Thus, MNEs are better able to lead the research and development domain than SMEs. This lack reflects the growing gap between small and large groups. The consequence of this gap is the low competitiveness of SMEs [Omri, 2020]. This sounds like the alarm

TABLE 1.1: SMEs characteristics based on their comparison with MNEs.

#	Attribute	Cluster	Big companies	SMEs
1	Technology	Resource	Developed	Medium
2	Finance	Resource	Available	Limited
3	Workers	Resource	Skilled	Polyvalent
4	Communication	Organization	Formal	Informal
5	Owners	Organization	Multi-national	Family
6	Decision	Organization	Informed	Based on feeling
7	Collaboration	Organization	Developed	Medium
8	Products	Method	Multiple	Specialized
9	Certification	Method	Exist	Rare

bells on the state of SMEs. This essential locomotive needs an overhaul more than ever to continue to shine and create wealth. In this context, the digitization process and the integration of Industry 4.0 technologies seem essential for the development and growth of SMEs [Moeuf, 2018, Omri et al., 2020]. However, many obstacles limit their transfer to Industry 4.0 [Mittal et al., 2018].

The next sections present the state of the art of industry 4.0 implementation within SMEs while discussing the identified barriers and challenges.

1.4 Industry 4.0

The emergence of several digital technologies has characterized the rise of the 21st century. These technologies have invaded our daily life, and the industrial field is not immune to this invasion. In this context, the term Industry 4.0 has been introduced as a concept for the implementation of these new technologies in the industrial field [Moeuf et al., 2018]. However, there are many other definitions for this concept.

1.4.1 Industry 4.0 definitions

Industry 4.0 refers to the 4^{th} industrial revolution, which aims to connect all objects and stakeholders along the factory's value chain. New technologies such as IoT [Dijkman et al., 2015], massive data analysis [Babiceanu and Seker, 2016], or Cloud Computing

[Xu, 2012] have recently entered industrial companies and promote the emergence of new concepts of process management, new services, and new products [Moeuf et al., 2017]. Consequently, several initiatives have emerged to manage these technologies. As an example of these initiatives, we cite "smart manufacturing" in the United States, "internet+" in China, "industry of the future" in France, and "industry 4.0" in Germany. The term "industry 4.0" seems to be gaining ground internationally [Moeuf, 2018]. For that, we propose to adopt this term throughout this manuscript.

The Industry 4.0 concept was introduced at the Hannover Fair in 2011 following a discussion between industry representatives, researchers, trade unions, and the state. The German telecommunications association BITKOM found in 2013 more than 100 different definitions of the Industry 4.0 concept [Bidet-Mayer and Ciet, 2016]. Thus, it still difficult to give a consensus definition of Industry 4.0. In [Sahal et al., 2020], Sahal et al. define the concept of industry 4.0 as a general framework that enables enterprises with new elements of tactical intelligence using new technologies such as the Internet of Things (IoT) and big data. In the same context, [Trappey et al., 2017] define Industry 4.0 as a general concept to make manufacturing smarter using techniques and technologies such as the Internet of Things, cloud computing, and big data. Moreover, industry 4.0 can be defined as "a new approach for controlling production processes by providing real-time synchronization of flows and by enabling the unitary and customized fabrication of products" [Kohler and Weisz, 2016]. For [Schumacher et al., 2016], Industry 4.0 refers to recent technological advances where the internet and technologies (such as embedded systems for example) serve as the basic support for the integration of physical objects, human actors, smart machines, product lines and processes across organizational boundaries to form a new type of smart grid and agile value chains.

Among all the existing definitions, we, therefore, select the one adopted by Moeuf et al. [Moeuf, 2018]: "Industry 4.0 is an approach to industrial management aimed at real-time synchronization of flows and the unitary and personalized manufacturing of products at the request of customers".

The objective of the Industry 4.0 concept is to make production methods more intelligent through the networking of machines and people [Moeuf, 2018]. Manufacturers can make strides toward Industry 4.0 through three pillars of thought: digitization, sensorization and optimization [Deng, 2021].

- **Digitization:** Before sensorization and optimization can occur, existing operations must be digitized to provide visibility over everything as it happens in real-time. This allows fault detection in real-time. Digitizing workflows allows labor and products to be monitored and actioned on immediately [Toth et al., 2018].

  Delays cost time and money, and digitization helps eliminate this by providing immediate alerts and notifications when operations go contrary to schedule. It also creates a treasure trove of data that can be used to optimize operations. It

allows the building of accurate cost models and enables manufacturers to focus on areas of inefficiencies and productivity improvement.

- **Sensorization:** The first step towards machine inter-connectivity is sensorization. IoT is leading this charge, which eliminates human monitoring and frees up resources for more critical areas. Sensors are cost-effective ways to measure variables such as temperature, moisture, air quality, motion, and vibration, among others. This enables equipment to auto-detect issues, which leads to auto-triggers and auto-configuration from a software and hardware perspective. For example, the implementation of non-contact temperature sensors allows for the auto-adjustment of roller speeds when gluing two cardboard pieces together at a cardboard manufacturing facility. This data can then be utilized to improve the bonding conditions and durability of the final product.

- **Optimization:** Admits all the data collected from digitization, sensorization, and integration. Transitioning this data into finding the hidden gems is the next priority. Optimization of manufacturing data comes through analytics, simulation, predictive and preventive maintenance, etc. Ultimately, the goal is to reduce costs and improve quality.

  The improvement and automation of technology don't require the complete removal of humans from the process; it's quite the contrary. Unique facilities require unique analytics and toolsets. This creates opportunities for industrial engineers, data scientists, manufacturing engineers, and statisticians and needs a new breed of data-driven manufacturing experts to derive improvements and optimization.

As mentioned above, the digitization process and the integration of Industry 4.0 technologies seem essential for SMEs' development and growth. However, many obstacles limit their transfer to Industry 4.0 [Mittal et al., 2018]. The next section presents the state of the art of Industry 4.0 implementation within SMEs while focusing on the limits and challenges.

1.4.2 Industry 4.0 implementation in SMEs

SMEs sustainability depends on their ability to satisfy their customers effectively [Li et al., 2016]. However, they still cannot benefit from the industry 4.0 technologies because they suffer from many problems limiting their adoption of such technologies [Omri et al., 2020]. In this context, many studies have been carried out in recent years to address this problem [Masood and Sonntag, 2020].

Figure 1.5 shows the annual evolution of publications dealing with Industry 4.0 and SMEs. In this section, we propose to review these publications while focusing on the advances in this domain and identifying barriers that limit the implementation of Industry 4.0 in SMEs.

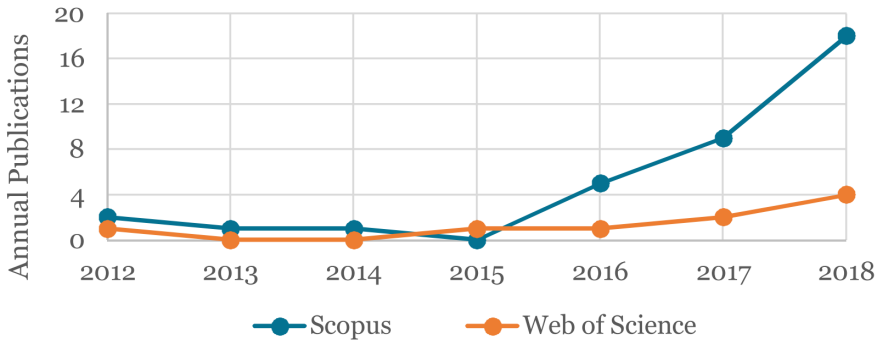


FIGURE 1.5: Annual publications dealing with Industry 4.0 and SMEs [Masood and Sonntag, 2020].

State of the art

The Industry 4.0 implementation in SMEs is a new topic that has gained more interesting in recent years [Masood and Sonntag, 2020]. One of the earliest works in this field was proposed by Würtz, and Kölmel [Würtz and Kölmel, 2012] where the authors pointed out the potential problems of implementing smart factories in small businesses. This study has inspired many subsequent works that have developed to analyze these problems and to propose adequate solutions to them [Rauch et al., 2018] [Sevinc et al., 2018] [Moeuf, 2018] [Moeuf et al., 2017] [Moeuf et al., 2018] [Mittal et al., 2018].

One of the first methodologies to implement Industry 4.0 was proposed by Wang et al. [Wang et al., 2016]. This methodology is based on five steps: preparation, analysis, idea generation, valuation, and implementation. However, these steps are disconnected from SMEs' reality and characteristics which limit their performance in real-world applications [Masood and Sonntag, 2020]. In the same context, the authors of [Wank et al., 2016] have proposed a similar conceptual methodology without taking into account the SMEs' specificities. Contrary to these studies, a real case study on three SMEs was proposed in [Jung and Jin, 2018]. The results were encouraging, but the authors stated that limited financial resources restrict the implementation of new technologies in SMEs. Therefore, the authors propose to overcome this problem by using low-level implementation. In the same context, the authors of [Pérez et al., 2018] highlight the issues linked to SMEs' acceptance for several technical reasons. To overcome this problem, many studies have been conducted to improve the awareness of the latest technologies in SMEs [Masood and Sonntag, 2020].

In [Kusiak, 2018], the authors report the benefits of implementing Industry 4.0 technologies: cost reduction, improvements in quality, efficiency, flexibility and productivity, and competitive advantage. However, SMEs can still not benefit from these technologies due to many barriers, as noted above. The following paragraph details these barriers to propose a new SME-oriented approach for Industry 4.0 implementation.

Barriers limiting the Industry 4.0 adoption in SMEs

Industry 4.0 is a set of cutting-edge technologies requiring specific resources that make its implementation within SMEs a problematic task. In [Omri et al., 2019], the authors review the SMEs domain and propose to classify the constraints that limit the development of such technologies within SMEs into two main classes: Resources-based constraints and Organization-based constraints. Moreover, Müller et al. [Müller and Voigt, 2017] conduct a study on German SMEs. They affirm that standardization, personnel resources, financial resources, and a belief in digitization are unique constraints for SMEs to integrate advanced technologies. In [Mittal et al., 2018], the authors claim that three main limitations hamper industry 4.0 implementation in SMEs: financial restriction, knowledge limitation, technology limitation. Based on an empirical study of 37 SMEs in Italy, Thailand, Austria, and the USA, Orzes et al. [Orzes et al., 2018] propose to group the significant barriers of Industry 4.0 implementation in SMEs in six main groups: economic and financial, cultural, competence and resources, legal, technical and implementation process. Müller et al. [Müller and Voigt, 2017] deployed the design interaction strategies for the introduction of Industry 4.0 in German SMEs and interviewed 68 experts, including 41 chief executive officers (CEOs) in firms dealing in mechanical and plant engineering, electrical engineering, and automotive suppliers, and concluded that standardization, personnel resources, financial resources, and a belief on digitization are unique constraints for SMEs. All these limitations are in line with the SMEs' characteristics discussed in Section 1.3.3. We propose to study the restrictions that limit the implementation of Industry 4.0 within SMEs: Human, Organizational, and Resources factors.

Resources factors: As mentioned earlier, SMEs are characterized by limited financial resources, which impact their ability to acquire advanced technological resources. In this context, Mittal et al. [Mittal et al., 2018] conclude that SMEs are financially constrained, which affects their ability to adopt advanced manufacturing technologies. Moreover, SMEs ' limited resources have influenced the research and development field among small enterprises [Julien and Ramangalahy, 2003a]. Based on an empirical analysis of 632 SMEs, the authors of [Kennedy et al., 2003] assert that SMEs are not involved in the deployment of advanced manufacturing technologies derived from Industry 4.0.

Organizational factors: SMEs are characterized by central management, where the owners are involved in all the decision-making process from the strategic level to the operational level [Bridge and O'Neill, 2012]. Thus, the first challenge in implementing Industry 4.0 in an SME is to convince the manager of these technologies' effectiveness. Moreover, a study of 600 Australian manufacturing SMEs proposed in [Terziovski, 2010] has shown that SMEs lack the innovation culture and strategy to succeed. Thus, the Industry 4.0 concept seems to be stranger to their environment. Also, communication in SMEs is generally informal and very close between workers [Durst and Runar Edvardsson, 2012]. Hence, information that concerns the manufacturing process is not

documented since it is kept in the mind of the manager and key workers [Omri et al., 2019]. Nevertheless, a reliable digitization process inside companies requires a certain level of documentation that guarantees no loss of knowledge [Yew Wong and Aspinwall, 2004].

Human factors: The human aspect is a crucial factor in the success or failure of Industry 4.0 [Masood and Sonntag, 2020]. In this context, Subrahmanya et al. [Subrahmanya, 2015] claim that unskilled workers join SMEs for a short period, and when they are fully equipped with the "required skills", they leave to join large enterprises. Consequently, advanced tasks cannot be supported by SMEs. Moreover, the received opportunities by SME employees are very different from those in MNEs [Mittal et al., 2018]. SME employees are known to be "jack-of-all-trades" [Bublitz and Noseleit, 2014], which prevent them from developing high skills in a particular domain [Dombrowski et al., 2010]. From an accessibility point of view, Industry 4.0 technologies are sufficiently complex to exceeds the user's capacity [Kothamasu et al., 2006]. This disadvantage creates a kind of mistrust between humans and these technologies, which limits their adoption within small organizations.

To sum up, the integration of Industry 4.0 technologies is required for the development and growth of SMEs [Moeuf, 2018, Omri et al., 2020]. However, many obstacles limit their transfer to Industry 4.0 [Mittal et al., 2018]. These obstacles revolve around the lack of a clear methodology for Industry 4.0 integration within SMEs. In this context, the PHM concept seems adequate for Industry 4.0 integration within SMEs [Omri et al., 2020]. PHM is an emerging concept in the industrial domain that brings together several disciplines for performance improvement. The interest of the PHM lies in the fact that it is a concept that starts from data through analysis and ends with a decision for performance improvement. Thus, we propose in the following sections to adopt the PHM concept as a solution to facilitate Industry 4.0 integration within SMEs. However, the PHM concept is not widely applied in the SMEs domain [Omri et al., 2019]. Thus, additional efforts are required to adopt this concept in SME cases, especially concerning the available data quality and cost.

1.5 PHM concept as key-enabler to implement Industry 4.0 in SME

The increased amount of data in the industrial field requires appropriate treatment to meet the challenge of zero defect manufacturing [Wang, 2013]. In this context, data-driven PHM of industrial systems has attracted the attention of researchers and industrialists during the last decade [Koulali et al., 2018]. Their works concern many fields such as the manufacturing, energy, and transportation industries [Kwon et al., 2016]. This section reviews these studies while presenting the conditions for implementing a data-driven PHM in the industrial domain.

1.5.1 The PHM paradigm

Prognostics and health management is a science that studies the health state of equipment and predicts its future evolution [Omri et al., 2019]. This concept allows better control of the systems and implementing suitable maintenance strategies [Pecht, 2009]. In [Omri et al., 2019], the authors define PHM as "a set of tools that can be used in cascade or separately to monitor the health state of a system, predict its future evolution and/or optimize decisions". In [Pecht, 2010], the authors affirm that PHM can be implemented using model-based or data-driven approaches. The first approach consists of building analytical models directly related to the physical processes that influence the health state of systems. Thus, a good comprehension of the physical process of components degradation and interaction is required. The second approach consists of using historical monitoring data to model the system's evolution until a failure occurs. In this case, the understanding of the system's physical process could not be necessary, but results only depend on the quality of historical data.

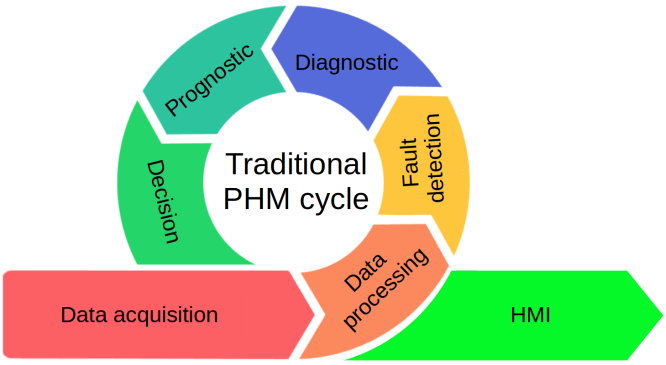


FIGURE 1.6: The traditional PHM cycle.

Traditionally, the PHM is decomposed into three main steps (observation, analysis, and decision), which are detailed in 7 steps from data acquisition to the Human Machine Interface (HMI) as shown in Figure 1.6. This configuration of the PHM process assumes that the studied system is already defined and that the necessary detection and analysis resources are available. Most of the existing PHM works do not cover how to determine the PHM system and assumes that the PHM will improve its performance [Adams et al., 2017]. However, there are many systems to be studied in the industrial domain with different impacts on the production system. Moreover, in SMEs' case, available resources are generally limited, and their allocation needs to be optimized to maximize the PHM study's benefit. In [Adams et al., 2017], the authors state that choosing the most suitable PHM project is a challenging activity. Thus, a clear estimation methodology is therefore needed. This methodology can be based on PHM indicators and their expected evolution after the application of the PHM. Moreover, the financial impact of the PHM depends on the available resources and data. Unless useful data (or the related resources) are available, a PHM project may be insignificant.

1.5.2 Towards predictive manufacturing: The evolution of PHM

As mentioned above, manufacturing systems have evolved throughout four industrial revolutions. This industrial evolution has been accompanied by a change in the manufacturing functions such as the maintenance one [Lee et al., 2014a]. This function was evolved from "repair work" to PHM (See Figure 1.7). For a long time, maintenance is considered as repair work where machines are not maintained except in the case of breakdown [Takata et al., 2004]. Since there is no available data for the machines, the prediction of the breakdowns was impossible. A little more time later, developments in the maintenance paradigm give birth to a new concept called time-based maintenance (TBM). TBM is based on the breakdowns historical of the machine to introduce preventive maintenance planning. However, this strategy ignores much information (since it does not exist) that can affect machine degradation. Because of these drawbacks and the evolution of the data technologies, the TBM was replaced by a more sophisticated maintenance method which is condition-based maintenance (CBM). The CBM is based on preventive actions taken after the apparition of failure symptoms [Xu et al., 2018]. Thus, CBM allows analyzing data coming from machines to avoid failure [Jardine et al., 2006, Lei et al., 2018]. Despite this significant evolution in the maintenance paradigm, many disadvantages have been entrusted to the CBM. The drawbacks concern the cost-effectiveness and the scope; in fact, this method becomes useless and expensive in the case of non-critical components. Moreover, CBM ignores many aspects such as safety, reliability, and the economic part. As a result, the CBM concept has evolved to give birth to the PHM discipline [Omri et al., 2019]. PHM goes beyond CBM scope and integrate many aspects such as logistics, security, reliability, mission criticality and cost-effectiveness [Vogl et al., 2019].

1.5.3 The evolution of data-driven industrial PHM Standards

PHM process involves different domains and applications, making it challenging to develop a general approach for PHM installation in enterprises. In this context, Vogl et al. [Vogl et al., 2014], assert that until now, there was no consistent guide for conducting a PHM study, and even the standardization of specific methods for PHM was ineffective since each application had its requirements. As the PHM paradigm has evolved from a maintenance function to a more global framework, documentation in this field has also evolved from a general maintenance context (FD X60250 standard) to a closer PHM context (ISO 13374). Many international institutes and organizations have worked in this domain and proposed standards and guidelines for PHM process implementation. These organizations include the Air Transport Association (ATA), the international organization for standardization (ISO), the international electrotechnical commission (IEC), the society of automotive engineers (SAE), and the united states army (US Army) [Guillén et al., 2016].

During the first and second industrial revolutions, no standards were identified in the context of asset maintenance and management, and even the few existing docu-

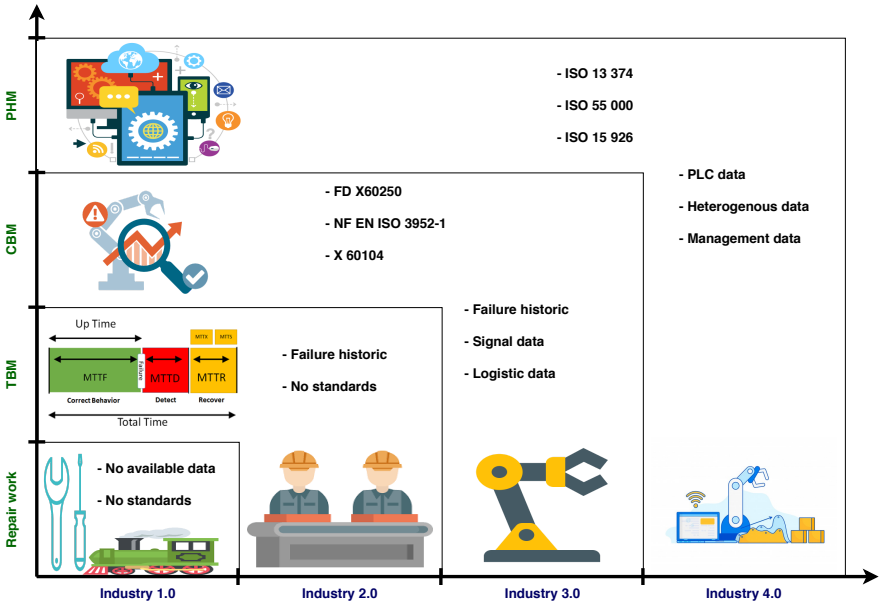


FIGURE 1.7: The evolution of maintenance paradigm within the industrial revolutions.

ments were only applied in the military domain. One of the first standards in this area is $X60104$ (1981), which defined maintenance contracts' requirements. Then, the ISO 3952-1 was published to introduce the symbols system and facilitate machine components documentation by modeling their interaction. Besides, the $FDX60250$ standard (1983) presents technical maintenance documentation for users and recommendations for its implementation. Later in the 2000s and with the emergence of data technologies, maintenance documentation has evolved to follow the technological revolutions. In this context, the $ISO55000$ standard has introduced the concept of asset management and widen the maintenance scope to integrate related activities such as the planning, design, implementation, and review of asset management activities. $ISO15926$ was published to represent process plant life-cycle information via a data model with a consistent context for data definitions. In the same context of data management standards, the ISO 13374 series (Condition monitoring and diagnostics of machines) was introduced. This standard is presented in 4 parts:

- ISO 13374-1 aims to provide the basic requirements for open software specifications to facilitate data transfer among various condition monitoring software, regardless of platform or hardware protocols [ISO, 2003].
- The ISO 13374-2 goes beyond the data transformation and provides requirements for a reference information model and a reference processing model for an open Condition Monitoring and Diagnostics (CM&D) architecture [ISO, 2007].

- The third part of the ISO 13374 standard (ISO 13374-3 [ISO, 2012]) defines the communication requirements for any open CM&D systems to aid the interoperability of such systems [ISO, 2012].
- The ISO 13374-4 [ISO, 2015] details the requirements for the presentation of information for technical analysis and decision support in an open architecture for condition monitoring and diagnostics.

To facilitate the uses of these standards, the Machinery Information Management Open Systems Alliance (MIMOSA) publishes an open CMD information specification known as the MIMOSA Open Systems Architecture for Enterprise Application Integration (OSA-EAI) [Mathew et al., 2006], which is free for download and compliant with the requirements outlined in ISO 13374-1 and ISO 13374-2. Based on the feedback of this standard, MIMOSA has elaborated another documentation (MIMOSA Open Systems Architecture for Condition Based Maintenance (OSA-CBM)), which is the most used in both the research and industrial domains [Thurston and Lebold, 2001].

1.5.4 Industrial PHM applications

The PHM concept has been widely applied in the industrial domain [Koulali et al., 2018]. In this context, Toshiba collaborates with NEC to develop an IoT-based PHM system. Thus, data are collected from Toshiba's devices and saved in data centers managed by NEC. Then the Toshiba maintenance team analyzes the data to respond to the customer's requests [Kwon et al., 2016]. A similar collaboration has been developed between Nidec and IBM. PHM services are proposed by IBMs based on the collected data from Nidec's machine [Kwon et al., 2016]. In the automotive industry, cars from General Motors, Tesla, BMW, and other manufacturers are equipped with application programming interfaces (APIs). The APIs allow applications built by third parties to use the collected data. This enables the development of applications for IoT-based PHM that add value by increasing connectivity, availability, and safety [Kwon et al., 2016]. The construction and mining industries benefit from developed connectivity technologies to control their equipment since working sites are generally isolated. In this context, Komatsu [Kwon et al., 2016] developed a data-driven PHM module to monitor and diagnose their construction equipment via satellite communications. In [Gao et al., 2018], the authors assume that PHM is a multidisciplinary activity that requires significant time and effort, making it very expensive. Feldman et al. [Feldman et al., 2008] apply the PHM concept to an electronics Line-Replaceable Unit (LRU) in the Boeing 737 aircraft. The results are obtained for 300 flights and shown that the PHM cost for each LRU is 700 $ (value is in 2008 U.S. dollars). These studies prove that PHM implementation requires a significant investment, which seems very expensive for an SME.

1.5.5 Research issues and assumptions

Based on the previous literature review, the advantages of Industry 4.0 are identified. However, it is still challenging to implement Industry 4.0 within SMEs since there is no

clear methodology. Thus, many authors have proposed a set of research gaps in this area [Mittal et al., 2018] [Masood and Sonntag, 2020].

In [Mittal et al., 2018], the authors claim that most established Industry 4.0 models, frameworks, and toolkits are developed for, or by, large MNEs, limiting their adoption by SMEs as they do not take into account the peculiarities of small companies. Moreover, the authors pointed out that there is no generic methodology to evaluate an SME's readiness to adopt Industry 4.0 technologies. Based on these research issues, Masood et al. [Masood and Sonntag, 2020] proposed a set of research questions that can be summarized in three points:

- Challenges related to the adoption of Industry 4.0 by SMEs.
- Industry 4.0 key technologies and their benefits for SMEs.
- Empirical method for assessing the characteristics of SMEs and the benefits of Industry 4.0 regarding these characteristics.

Based on these research issues, we propose to develop an SME-oriented methodology for the implementation of Industry 4.0. To do this, we propose to adopt a prognostics and health management (PHM) concept as a key-enabler to implement Industry 4.0 in SMEs. Recall that the PHM is not the only solution but, it is only a concept that we adopt in this Work. This choice is justified by the fact that PHM is widely applied in the industrial domain with a clear methodology [Vogl et al., 2014]. Moreover, the PHM approach is similar in its principles to the three Industry 4.0 pillars discussed in Section 1.4.

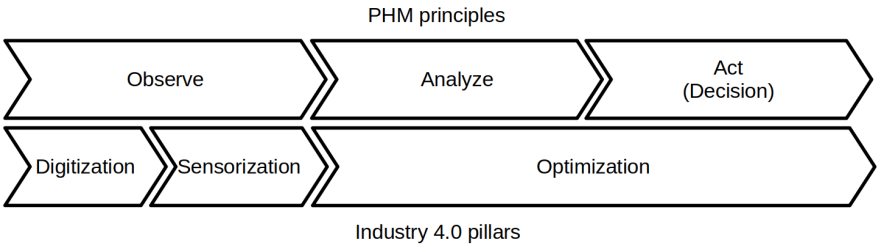


FIGURE 1.8: PHM - Industry 4.0 analogy proposal.

Figure 1.8 shows an analogy between the PHM principles and the Industry 4.0 pillars. The digitization and sensorization pillars from the Industry 4.0 can be grouped in the PHM approach's observation step. Indeed, The analysis and act steps can be grouped in the optimization pillar.

Generally, the development and deployment of PHM within an industrial organization is a very complex task. In [Guillén et al., 2016], Guillen et al. affirm that there is a gap in PHM documentation. Establishing general methodological approaches to guide the design and implementation of the PHM process is hence needed. Moreover, up to our

knowledge, there are no documentations dedicated to SMEs. In [Kennedy et al., 2003], the authors concluded that SMEs are not involved in deploying advanced manufacturing technologies. Thus, SME constraints are studied and discussed in the next section to propose a generic PHM implementation strategy.

The following section will be devoted to present the PHM concept while detailing the proposed SME-oriented PHM approach to implement Industry 4.0 within SMEs.

1.6 An adapted PHM approach for Industry 4.0 implementation within SMEs

Considering the previously detailed constraints, the best appropriate way to implement Industry 4.0 technologies within SMEs is to follow a PHM process adapted with SMEs' characteristics [Omri et al., 2020]. Thus the digitization phase should start with existing data [Omri et al., 2019]. Before collecting new data, it is necessary to digitize the current data. Since SMEs do not have many resources (financial resources in particular) to install sophisticated data acquisition devices. It is recommended to use simple acquisition systems and use non-expensive storage solutions. Thus, smartphones and tablets are the simplest solutions when they are coupled with existing and free mobile applications. Free cloud solutions could be used to ensure real-time acquisition, but this solution calls into question the confidentiality and security of the collected data [Omri et al., 2020]. Once the company's ordinary data is digitized, potential valorization applications are discussed. A matrix that links each data group to the valorization application with its associated benefit can help to set priorities for the data analysis phase. One should note that the SME world is not accustomed to sophisticated processes such as data-driven PHM, so it is better to think about working quickly with existing data to provide useful results and prove the project's feasibility quickly. This step can convince the managers, implicate workers, and introduce Industry 4.0 culture into the company at the same time.

We propose a set of best practices that should be followed to successfully integrate Industry 4.0 within SMEs using an SME-oriented PHM process (See Figure 1.9). The main phases of this approach are detailed in the following.

1.6.1 Data inventory

Data inventory is a deep analysis of the circulating data around the manufacturing process. More particularly, data inventory is a quest to collect information about the existing data [Omri et al., 2020]. These information concern different elements present in Table 1.2.

In addition to the process of collecting information about the data, data inventory also aims to regroup data around potential PHM projects and identify the missed data to accomplish these projects. To do this, a list of necessary attributes for each system can be defined. This list depends on the nature of the studied system. In [Cheng et al.,

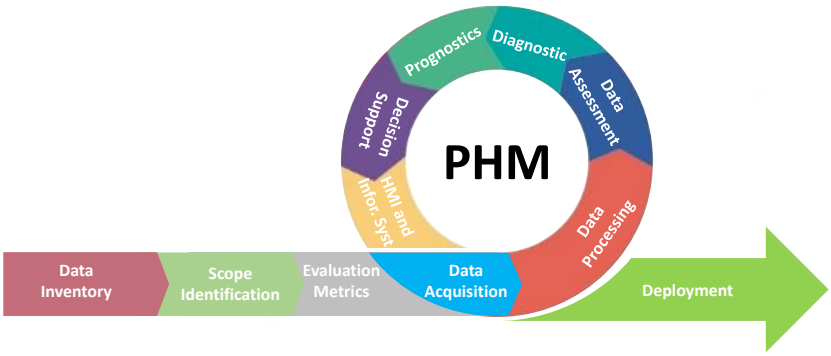


FIGURE 1.9: The extended PHM cycle [Omri et al., 2020].

TABLE 1.2: Attributes of a Data Inventory Quest.

Information	**Description**
Title	Dataset name
Features	Dataset attributes
Purpose	Data creation aim
Type	Text, images, numbers
The owner of the data	Production team, Marketing team
Location of the data	ERP, Server
The volume of the data	1 GB, 1TB
The format of the data	Papers, CSV files
Data transfer	Data usgae per team
Update frequency	1 hours, 1 day, 1 week
Restrictions	Confidentiality, Accessibility

2010], the author proposes a set of variables that can be collected to make a PHM study on eight different domains (see [Cheng et al., 2010] for more details). This list can help define the needed variables for each system and then compare them with the existing ones to identify the missed variables.

1.6.2 Scope identification

Since available resources in SMEs are limited, it is mandatory to limit a PHM study scope and focus only on the relevant projects [Omri et al., 2020]. Thus, the data inventory step's potential projects should be ranked and select the most relevant ones. Projects relevance depend on the following points:

- The available digital data.
- The data to be collected or digitized.

- PHM profitability.

To do, some techniques are proposed in the literature. As an example of project raking techniques, we can cite the Analytical Hierarchical Process (AHP) [Saaty, 1980] and the Data Envelopment Analysis (DEA) [Cooper et al., 2004]. In the PHM case, some works are done to optimize the sensors selection [Xu et al., 2015]. The selection process is generally converted into an optimization problem that can be solved by traditional techniques such as linear programming. The common point between these techniques is the assignment of an importance order to each characteristic to prioritize them and calculate a global score for each system. This score is used to rank the available equipment, select the relevant PHM equipment applicability, and guide the investment plan to maximize the study's benefits.

1.6.3 PHM metrics

One of the most critical steps in a PHM study is evaluating the whole approach [Omri et al., 2020]. To do, a set of metrics is needed to assess the project's objectives, either for big or small companies. Also, metrics are required to describe better the performance of the used technologies to satisfy these objectives. In [Luna, 2009], the author proposes a set of PHM metrics concerning the different PHM themes and theirs benefits. We here propose to classify the PHM metrics in relation to the company's characteristics, the objectives, the collected data, and the used PHM techniques.

TABLE 1.3: Example of PHM metrics in SME [Omri et al., 2020].

Company	**Objectives**	**PHM tools**	**Data quality**
R & D budget	Reduce maintenance frequency	Accuracy	Completeness
%of skilled workers	Improve products quality	False alarm	Time to value
Documentation level	Optimize resources allocation	Prediction lead time	Volume

Table 1.3 shows an example of PHM metrics used in the different steps of a PHM study. Accordingly to the fixed objectives, a set of performance of the used techniques can be identified. Moreover and based on this performance, the data required data quality can be fixed. In a real case study, it is sometimes challenging to ensure the required data quality, necessitating a modification in the initial objectives [Omri et al., 2019]. PHM metrics are common to all companies' sizes, but one should think about specific metrics for SMEs in line with their limits. In this context, the percentage of skilled workers, the research and development (R & D) budget, and the documentation level can be used as PHM metrics in SMEs' case.

1.6.4 Data acquisition, analysis and valorization

Industry 4.0 is considered as a data revolution [Omri et al., 2020]. Thus, data acquisition, analysis, and valorization are critical tasks. From a PHM point of view, these tasks correspond to the traditional PHM cycle. However, they need to be modified and adapted to the SME characteristics. In this context, the weak data quality SME is the main problem that limits these tasks. For that, a major part of this work is dedicated to address this problem.

1.6.5 HMI and deployment

The previously detailed steps are developed to integrate Industry 4.0 within SMEs successfully. However, it remains another challenge: the facilitation of these tasks for non-specialized users [Omri et al., 2020]. The deployment of such technologies within SMEs is limited by the "jack-of-all-trades" character of SME operators. We propose to formalize the proposed SME-oriented PHM approach in ergonomic graphical interfaces where the user can easily accomplish these tasks.

To sum up, the proposed PHM strategy extends the PHM cycle (See Figure 1.9). This process is indeed very long, but in our opinion, it is the most adapted one to implement Industry 4.0 within SMEs with few resources, as mentioned previously. The proposed SME-oriented PHM approach is applied in the SCODER company, which is a French SME. The conducted literature review combined with this real-world application allows identifying the research gaps that need to be filled to implement Industry 4.0 within SMEs successfully. In the sequel, the results of this study are reported, and the identified challenges are discussed.

1.7 Implementation of the proposed approach in the SCODER case study

In this section, we present the different steps of a real application of the proposed approach. This application consists of a cutting and stamping line from the SCODER factory. Indeed, this line is exposed to several production anomalies that result in production stoppages or quality problems. The objective is to digitize and control this line to reduce these anomalies and ensure stable production in productivity and part quality. However, SCODER is an SME with limited resources, which means an additional effort to succeed in this process.

The following paragraphs detail the different steps of integrating Industry 4.0 technologies within this production line.

1.7.1 Digitization step

As shown in Figure 1.10, the first step of the Industry 4.0 implementation is the digitization step. To successfully conduct this phase in the case of an SME, the scope of the study should be carefully selected. For that, a data inventory query must be carried out before as detailed in the proposed approach.

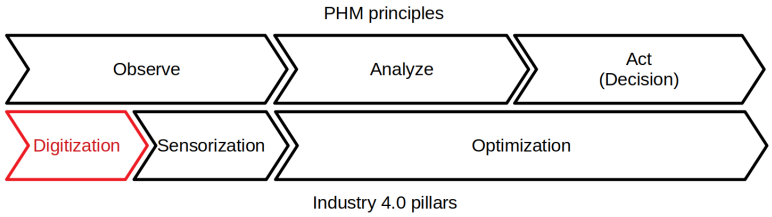


FIGURE 1.10: Industry 4.0 implementation in the SCODER company: the digitization step.

Data inventory: Different data which comes from heterogeneous sources are shared in the factory. These data are generally recorded in papers that make it difficult for their deployment in this study. As proposed in the strategy, the first step to conduct a PHM study within SMEs is to start with the data inventory. Figure 1.11 describes the physical and informational flows in the factory.

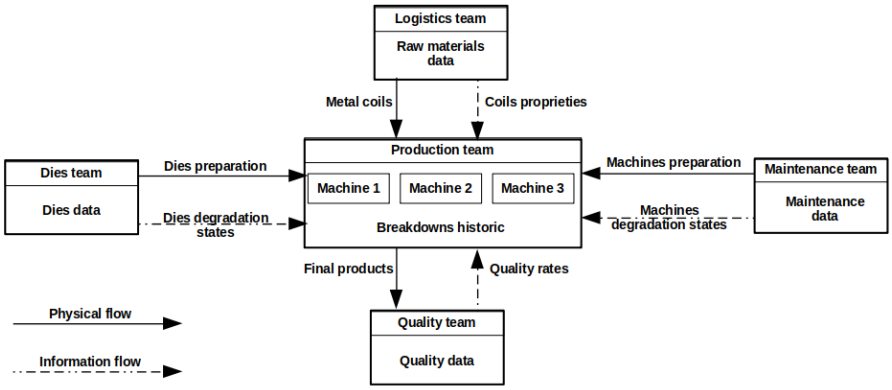


FIGURE 1.11: Data inventory in the SCODER case study.

Scope identification: The objective is to ensure stable production, but we can only study one machine with the available resources. A scope selection is first made according to the characteristics mentioned in Subsection 1.6.2. Then, these characteristics are used to affect a score empirically to each machine and rank them to select the most suitable one for a PHM study (See Figure 1.12). One should note that it could be possible to

formulate this score mathematically based on the conducted PHM projects' feed-backs. To take into account the importance level of each characteristic, a constant coefficient is affected to them. Then the final score is calculated as below:

$$Score = 0.2 \times available\ digital\ data + 0.3 \times data\ to\ be\ collected + 0.5 \times PHM\ profitability. \tag{1.1}$$

Figure 1.12 shows that the third machine is the most suitable to conduct a data-driven PHM study.

		Systems		
		Machine 1	Machine 2	Machine 3
Characteristics	available digital data	2	4	4
	data to be collected	2	3	4
	PHM profitability	4	4	5
Score		3	3.7	4.5

FIGURE 1.12: Machines ranking for potential data-driven PHM projects. Red color refers to a weak score, yellow color represents an acceptable score and green color indicates a good score.

The PHM project was initiated to ensure stable production by reducing machine failures and improving productivity. The production performance is affected by the used metal, the die, and the mechanical press. However, a study was conducted inside the factory showed that the used metal coil characteristics have the most critical impact on production. A PHM study is conducted from these existing data to determine an "Id card" for each sheet metal coil. This Id card represents the coil's characteristics, the caused press breakdowns, and the quality rate of the products fabricated from it. However, only the sheet metal characteristics are available without any indication about the quality rate, date, and time of each metal coil's use.

1.7.2 Sensorization step

As shown in Figure 1.13, the second step of the Industry 4.0 implementation is the sensorization step.

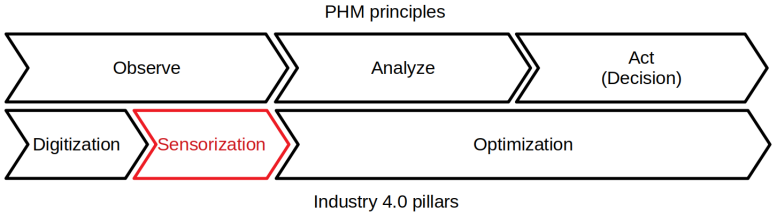


FIGURE 1.13: Industry 4.0 implementation in the SCODER company: the sensorization step.

This step aims to ensure the total connectivity of the studied machine. For that, a data acquisition system is installed on the SCODER's machine. This system consists of using a tablet at the beginning of the production line to scan the bar code of each coil to save its date of use, and another tablet at the end of the production line is used to collect the quality data (See Figure 1.14).

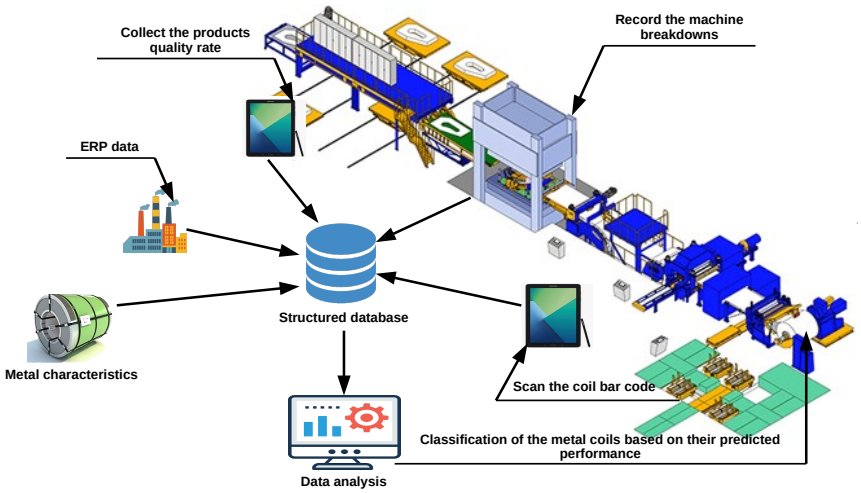


FIGURE 1.14: Details of the SCODER case study.

These data are coupled with the machine breakdown data to describe the production process. Data are structured and saved to be analyzed later using intelligent algorithms and extract knowledge from them.

1.7.3 Optimization step

As shown in Figure 1.13, the final step of the Industry 4.0 implementation is the optimization step. To do, the created dataset is used to extract knowledge and to optimize the production process.

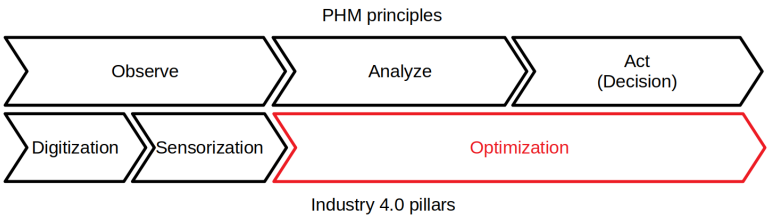


FIGURE 1.15: Industry 4.0 implementation in the SCODER company: the optimization step.

In this case study, the decision tree (DT) algorithm is used to classify the metal coils in different categories regarding their expected performance. The DT method is chosen

because it is an explainable machine learning tool, which means that workers know the built classification rules. Figure 1.16 shows the evolution of the accuracy rate in function of the volume of training data (since the real output of the n^{th} training iteration added to the training subset of the $(n+1)^{th}$ iteration). One can point out that after a few iterations, we can reach about 50% of accuracy. This low accuracy could be explained by the fact that the existing data doesn't contain information about the die's health state and the press. Moreover, the data volume is not sufficient.

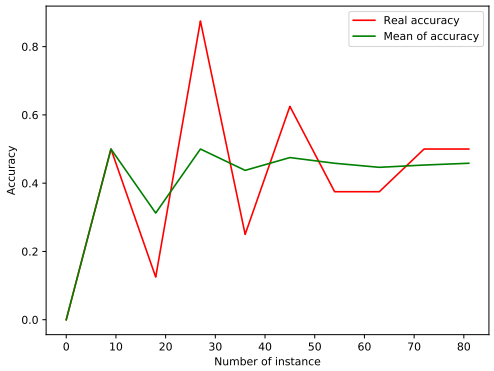


FIGURE 1.16: Evolution of the accuracy function of the number of instance used in the training phase. These results are obtained using the DT algorithm.

To sum up, these weak results can be explained by the bad quality of the available data (insufficient features, low volume, etc.). Thus, a reliable data quality study that respects SMEs' characteristics should be conducted. The next section discusses the different gaps that should be filled to implement Industry 4.0 within SMEs using a PHM approach.

1.7.4 Discussion

Recall that this work addresses the two most critical issues that limit Industry 4.0 within SMEs. These problems are the lack of adapted methodology and the identification of research gaps and their solutions. Based on this study, a PHM-based strategy is proposed to implement Industry 4.0 technologies in SMEs while focusing on the encountered problems. These problems concern especially the lack of resources and the specific SMEs' organization. Table 1.4 summarizes the main differences between big companies and SMEs for Industry 4.0 implementation as well as the proposed solutions for each challenge.

1.8 Conclusion

The SCODER company's motivation to install Industry 4.0 technologies as part of its digital transformation leads to express an industrial problem raised by the daily concerns of the workshops. This industrial problem concerns the challenge of installing Industry

TABLE 1.4: The proposed PHM-based strategy challenges and solutions.

Factor	Attributes	MNEs	SMEs	Proposed solutions
Finance	Budget	Available	Limited	Performance improvement at right costs
Technology	Infrastructure	Available	Limited	Scope identification
	Data quality	Medium	Medium	DQ assessment and improvement
Human	Skilled workers	High	Medium	Ergonomic graphical interfaces
	Resistance to change	Low	Medium	Automate the existing data projects
Organization	Documentation	Developed	Semi-developed	Knowledge capitalization process
	Objectives metrics	Detailed	Global	Standards PHM metrics

4.0 technologies within SMEs with limited financial resources. Thus, this double observation, i.e., on the one hand, towards the statement of the industrial problem and the related scientific questions, and on the other hand, towards the gap in the economic aspect of the Industry 4.0, are the main drivers of the thesis.

In this thesis, we propose to use a new SME-oriented PHM approach to implement Industry 4.0 within SMEs. This choice is justified by the fact that PHM is widely applied in the industrial domain with a clear methodology, and its principles are similar to the Industry 4.0 pillars. In this chapter, the general PHM methodology is modified to be adequate to SMEs characteristics based on an in-depth literature review and the expertise of our industrial partners in the SCODER company. Based on this collaboration, the different barriers and difficulties have been identified and discussed in Section 1.5.

In the rest of this work, we propose to address the data quality problem and its impact on the success of the Industry 4.0 integration within SMEs. For that, an in-depth study of the data quality in the industrial domain in general and in SME, in particular, is proposed. Moreover, a new approach is proposed to improve data quality and thus the PHM performance at the right cost. This approach is developed with respect to SME characteristics.

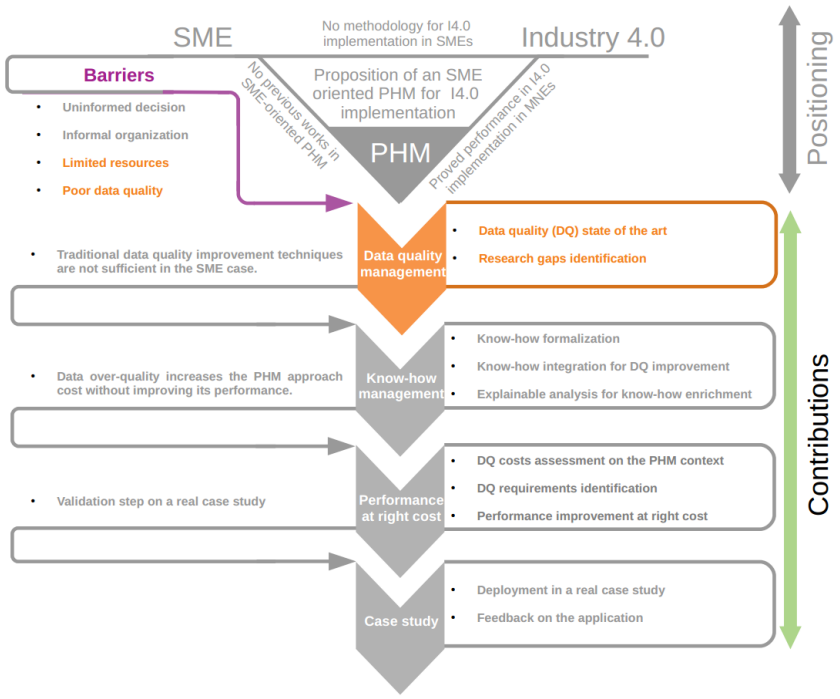
Chapter 2

Data quality and their improvement techniques: state of the art

"There is no shame in ignorance, only in denying it. By knowing what we do not know, we can take steps to remedy our lack of knowledge."

-Graham McNeill

Graphical abstract.



Contents

Contributions

This chapter proposes an overview of data quality management, covered by state of the art in data governance. Data quality dimensions and applications for this domain are also presented, with data quality and its definitions in the PHM context as the focal point of the approach. To improve the analysis results, notably by improving the used data quality, an overview of the adopted improvement techniques is proposed, addressing the most encountered problems in the PHM context. Based on this literature review, we identify the data quality management within SMEs as the central research gap, for which we propose to fill in the next chapter.

2.1 Introduction

Industry 4.0 is a data-based revolution that lies in digitization and sensorization to optimize performance. However, small and medium-sized enterprises (SMEs) still cannot benefit from this revolution due to many barriers discussed in the first chapter of this manuscript. To overcome these barriers, we propose to follow an SME-oriented prognostics and health management (PHM) approach to facilitate the integration of Industry 4.0 technologies within SMEs. The proposed method was applied to a real case study

where several research gaps have been identified and discussed in the previous chapter. This chapter deals with the first research gap, which concerns the formalization of the most encountered data quality problem in the industrial domain in general and in the SMEs in particular. For that, a state of the art on data quality and their improvement techniques is proposed in this chapter.

The remainder of this chapter is organized into seven sections. Section 2.2 presents data in the industrial domain while focusing on the evolution of their role during the various industrial revolutions. Moreover, the different industrial data sources are presented in this section. In Section 2.3, the data quality management problem is illustrated while identifying the main research gaps. Section 2.4 introduces the data quality (DQ) concept as a pillar of the data governance framework while detailing different data quality dimensions (DQDs). The most encountered data quality issues in the PHM domain are presented in Section 2.5. Section 2.6 gives a brief review of data quality improvement techniques that deal with these data quality issues. A synthesis of this state of the art is given in Section 2.7. Finally, conclusions are drawn in Section 2.8.

2.2 Industry 4.0: a data-based revolution

Data are real-world objects with storage, retrieval, and development capabilities, and they can be communicated over a network [Kong et al., 2020]. The process that deals with their valorization are called Data management. In [Sebastian-Coleman, 2012], the authors start from the definition of "management" which means "the process of dealing with or controlling ... [and] having responsibility for ...", and they claim that data management is more than "deal with "since it includes many tasks such as identifying resources to meet objectives, organizing these resources, defining and implementing a data management strategy to achieve a fixed set of objectives. For the Data Management Association (DAMA) [DAM, 2017], data management is "the business function that develops and executes plans, policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data". In the same context, another definition is proposed by Fisher in [Fisher, 2009], which considers data management as "a consistent methodology that ensures the deployment of timely and trusted data across the organization". In [Vogl et al., 2014], authors affirm that data management concerns data collection, processing, visualization, and storage. Hereafter, we are interested in the first two steps (data collection and processing) by detailing the industrial data sources and studying the impact of the quality of these data on the processing phase.

2.2.1 Historical perspectives and problem statement

As shown in Figure 1.7, data are always collected in the industrial domain. However, these data's nature and utility have evolved with the evolution of the manufacturing process. The new needs in terms of product quality and process optimization have forced companies to collect more data to increase the reliability and capability of their manufacturing processes [Omri et al., 2019]. This numerical transformation has been supported

by the advances in the Information Technologies (IT) [Tao et al., 2018]. These technologies have enabled enterprises to have more accurate information about their production process. Thus, the valorization of the industrial data has been evolved from absenteeism rate calculation to a more global PHM framework that allows optimization of the maintenance function, improving the production process, and reducing operational cost. This section proposes to review the evolution of industrial data collection throughout the various industrial revolutions. We propose to study the impact of this evolution on developing the maintenance function within industrial companies. Moreover, the change of documentation that standardizes the data management process for maintenance applications is addressed (See Figure 1.7).

The evolution of manufacturing data

For a long time, data is generated throughout the manufacturing process. The complexity of these data depends on the complexity of the industry and its degree of development. Usually, data are recorded with different technologies, from paper to big data hubs. The utility of these data depends on the needs of the companies. The role of data has evolved in line with the evolution of the industrial activities, which are characterized by four revolutions (from industry 1.0 to industry 4.0). For that purpose, we propose to study the manufacturing data in line with these revolutions.

The first industrial revolution (industry 1.0) The introduction of steam engines triggered this revolution into the production processes [Madsen et al., 2016]. According to Tao et al. [Tao et al., 2018], this revolution has no impact on the collection, storage, and analysis of data. Only a few data are managed manually by the workers through papers, but the rest is stored in human memory as know-how. These data usually relate to workers and serve minimal purposes. Thus, information such as assistance, productivity, and performance are collected to evaluate worker performance.

The second industrial revolution (industry 2.0) The second revolution is characterized by mass production. Electric machines are used with more sophisticated management principles (e.g., the Bessemer process, the Taylorism, etc.) [Madsen et al., 2016]. Unlike the first revolution, Industry 2.0 has imported many changes in data management. More data are recorded in more formal documents such as charts, and logbooks [Tao et al., 2018]. As a result, these data's utility has been expanded to reach other applications such as operation planning, quality control, and failure rate.

The third industrial revolution (industry 3.0) The third revolution was introduced thanks to the development introduced in the domain of computers and semiconductors in the 1960s [Schwab, 2017]. This revolution is characterized by the use of Logic Controllers, Computer Numerical Control (CNC) robotics, which gives birth to the concept of fully automated factories [Madsen et al., 2016]. As a result, data are collected automatically, saved in computers, and managed by information systems. From these

data revolution was born a set of information systems (e.g., ERP, MES, etc.) to manage the maintenance, production, supply chain, and financial data [Tao et al., 2018].

The fourth industrial revolution (industry 4.0) The fourth industrial revolution was triggered thanks to the emergence of the Internet of Things (IoT), Artificial Intelligence (AI), and Big Data analytics, which are integrated into the manufacturing process. All these technologies make it possible to collect and manage a massive quantity of data from several heterogeneous sources [Tao et al., 2018]. These data describe the product throughout its life-cycle [Li et al., 2015].

To sum up, the evolution of data collection, technologies, and utility are evolved in line with industries' development. Also, the data analysis techniques are evolved increasingly to valorize these data and extract knowledge from them. These techniques can be regrouped in a general framework such as the PHM [Guillén et al., 2016]. The evolution of this concept is discussed in the next section.

2.2.2 Industrial data sources

As mentioned above, the PHM implementation must take into account the existing data in the enterprise and propose adequate solutions to deal with the problems that characterize these data. Here, we propose to clarify the data architecture in the manufacturing organizations and detail the various data sources used as input in a data-driven PHM process. In this context, data can be extracted from different levels and sectors of production. These levels are usually represented in a pyramidal architecture where information flows from the bottom to the top of the pyramid, unlike the control flows that flow from top to bottom [Hoffmann et al., 2016]. Figure 2.1 represents the different data levels of an industrial company. This pyramid is decomposed into Enterprise Resource Planning (ERP), Manufacturing Execution Systems (MES), Control Level, and Device Level from the top to the down.

ERP systems are created as a solution to a classic problem in the industrial field. This problem involves treating the activities and transactions separately, without any link between them [Ross and Vitale, 2000]. For this, ERP systems provide a common systems platform that enhances data visibility. In [Madanhire and Mbohwa, 2016], ERP is defined as a method for planning and controlling the required resources to respond to customer orders. These tasks are satisfied using a software package that manages the data flowing in the enterprise. These data concern a global representation of the companies, including the different resources (human, machines, and raw materials). Also, the ERP database includes information about sales, historical production data, accounting, and production range [Madanhire and Mbohwa, 2016]. ERP systems are used for long-term activities planning without focusing on the shop schedule to accomplish these tasks [Hoffmann et al., 2016]. Unlike ERP systems, the MES software focuses on digitizing the production process to enable real-time control of the different activities [Coronado et al., 2018]. The MES provides information to optimize activities throughout the production

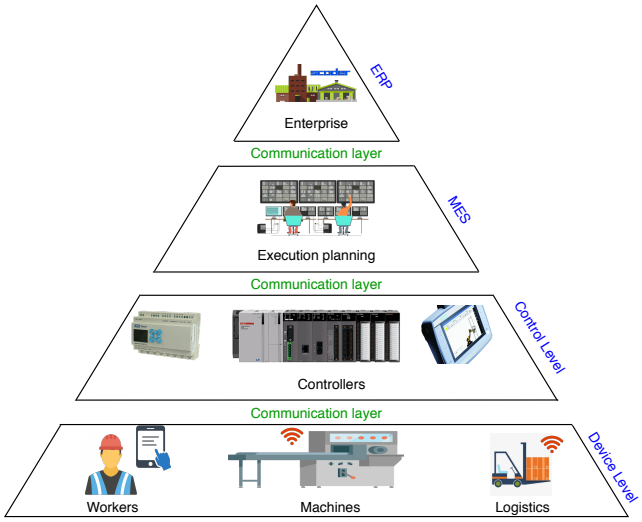


FIGURE 2.1: Automation pyramid in the industrial domain.

process. Using current and accurate data, an MES system guides, initiates, responds, and reports on workshop activities as they occur. [Saenz de Ugarte et al., 2009]. In other words, the MES manages finer data in terms of granularity. These data concerns manufacturing instructions, design engineering data, the status of resources, the progress of activities, and all events that occurred during the production activities. The control level is composed of all forms of computer or programmable cards that control the system state's evolution during its operational mode to detect and avoid failures. Also, these digital computers are responsible for controlling the industrial environment to reflect the workshop's atmosphere. The device-level is characterized by low-level devices that are represented by the machines or sensors. These devices generate data needed to perform process optimization or detect problems in the production flow.

2.3 The problem of data quality management

Many types of data can be collected from the previously detailed data sources, such as tabular, image and time series, etc. [Omri et al., 2020]. These data generally present some quality problems. However, data analysis results depend heavily on the quality of the input. In this context, Hyunseok [Oh et al., 2018] describes this phenomenon using a well-known proverb, "Garbage in, garbage out," which means that if the used data are of low quality, the poor results are unavoidable. Traditionally, data quality management follows a straightforward process where data quality is assessed, and its suitability for the application is evaluated via the obtained results [Omri et al., 2021]. This implies that the data acquisition step is carried out in advance without considering the fixed objectives.

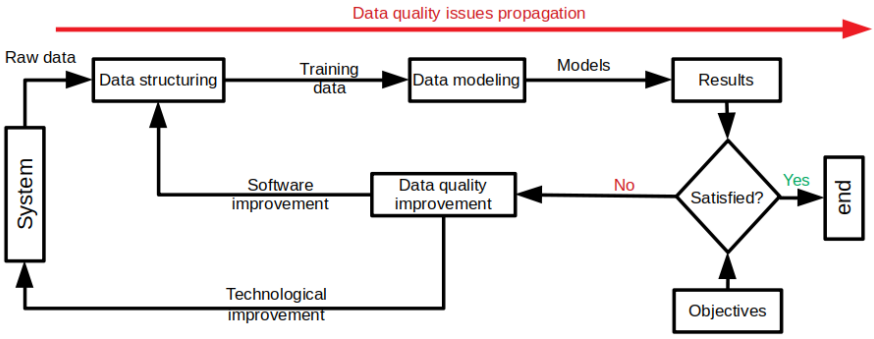


FIGURE 2.2: Traditional data management process.

As shown in Figure 2.2, the data quality impact propagates through the various analysis processes (data structuring and modeling) until it influences the results [Omri et al., 2020]. Thus, two cases can occur: either the obtained results correspond to the set objectives, which means that the used data are of good quality, or the objectives are not yet met. In this second case, two solutions are proposed in the literature [Omri et al., 2021]:

- **Software data quality improvement:** in this case, data quality improvement algorithms are used to enhance, such as imputing the missing data or reducing the noise.
- **Technological data quality improvement:** in this case, more sophisticated sensors are installed in the system to enhance the collected data quality.

Recall that this work concerns SMEs, which are generally limited in terms of financial resources. Therefore, technological solutions to improve data quality are not considered because they can generate high costs. Thus, we propose here to review the data quality improvement algorithms concerning the most encountered data quality problem in the industrial domain in general and in the SMEs in particular. In this context and concerning the SMEs' data problems, the authors in [Omri et al., 2019] report that the well-known issues are the missing data, the manually recorded data, the small volume of data, and the irrelevant data. Missing data refers to incomplete elements in a database. These missing data are due to a problem in the acquisition system or a difference in acquisition frequency. For manually recorded data, this problem applies to all businesses regardless of size. Still, it should be noted that this problem is more related to SMEs because they do not have the technology to digitize their data. The following sections are dedicated to detail the data quality issues while focusing on those that affect the SMEs' data.

2.4 Data quality in the industrial context

This section discusses the state of knowledge primarily related to data quality dimensions (DQDs). The link between data governance and DQ is first elaborated. Further, DQDs are defined and discussed in a general context and the PHM context.

2.4.1 Data Governance and Data Quality

Increasingly, extracting knowledge from data has become an essential task in organizations for performance improvements [Omri et al., 2020]. Thus, organizations are gradually implementing data governance policies [Juddoo et al., 2018]. In [Russom, 2008], the authors affirm that data governance should be an organization-wide concern and denounce its underestimated in many organizations. In this context, the DG framework of the data governance institute (DGI) [Institute, 2021] defines DG as "the exercise of decision making and authority for data-related matters" [Juddoo et al., 2018], and the main benefits for organizations adopting DG are (1) revenue and value increasing, (2) cost and complexity management, and (3) compliance, security, and risk control related to privacy.

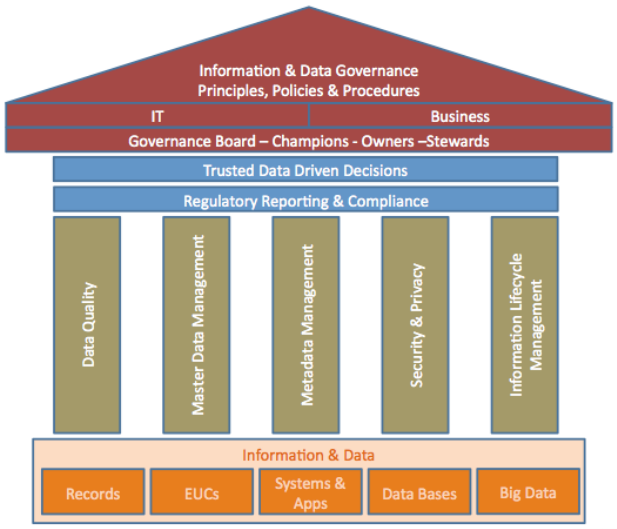


FIGURE 2.3: The DGI data governance framework [Institute, 2021].

Figure 2.3 shows the main pillars of the DGI data governance framework. The first pillar in this framework is data quality (DQ).

Data quality (DQ) has been the subject of many research works where several definitions were proposed to characterize this concept. The ISO 8000-8:2015 standard [ISO/IEC, 2015] describes fundamental concepts of information and data quality and how these concepts apply to quality management processes and quality management systems. In [Zaveri et al., 2016], Zaveri et al. assume that data quality problem refers to

a set of issues that can affect the potentiality of the applications that use the data. The authors in [Omri et al., 2020] affirm that most of these definitions link data quality to a set of requirements to satisfy. The ISO/IEC 25012 standard [ISO/IEC, 2008] definition assumes that high data quality is "the degree to which a set of characteristics of data fulfill requirements". Indeed, authors in [Sebastian-Coleman, 2012] define it as "data that is fit for use by data consumers". Data quality is usually defined according to a set of requirements that should be accomplished. We here adopt the data quality definition proposed in [Omri et al., 2020] and assumes "high-quality data as all data with a minimum level of quality that guarantees the satisfaction of objectives set by the owner".

As previously stated, data quality is a multidimensional issue that is widely studied in the literature. Thus, a set of Data Quality Dimensions (DQD) is defined to characterize the data requirements [McGilvray, 2008, Sidi et al., 2012]. The next section presents the essential data quality dimensions discussed in the literature while giving a brief interpretation of them.

2.4.2 Data Quality Dimensions

To ensure the best data quality understanding, the first step consists of studying the data quality dimensions (DQD) [Batini et al., 2009]. DQDs are defined as the means of expressing the notion of data quality such as consistency, accuracy, completeness, and timeliness [Juddoo et al., 2018]. This section presents the essential data quality dimensions discussed in the literature while giving a brief interpretation of them.

Many DQ dimensions were proposed and discussed in the literature, but until now, there is no consensus on the essential DQ dimensions for DQ evaluation [Sebastian-Coleman, 2012]. However, there is a shortlist of DQ dimensions, which are the most mentioned and discussed in the literature. Wang and Strong in [Wang and Strong, 1996] have presented one of the first approaches for DQ dimensions identification and present many DQ dimensions, which are reduced in a first step in 20 dimensions and then in only 15 dimensions. This list of dimensions is displayed in Table 2.1.

According to Redman, [Redman, 1997], this list can be reduced more to a concise list that contains only: Accuracy, Completeness, Timeliness, and Consistency.

In a general way, DQDs express the notion of data quality, and as shown above, data quality depends on the users' needs. For that, many authors in the literature assessed that data quality could be task-independent and, therefore, not restrained by the application's context, while others are task-dependent [Pipino et al., 2002b]. Regarding the work proposed by Wang and Strong [Wang and Strong, 1996], DQDs can be regrouped in 4 main classes:

TABLE 2.1: Most cited data quality dimensions [Pipino et al., 2002a].

Dimension	Definition
Accessibility	Extent to which data is available or easily and quickly retrievable
Appropriate amount of data	Extent to which volume of data is appropriate for the task at hand
Believability	Extent to which data is regarded as true and credible
Completeness	Extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Consistent representation	Extent to which data is presented in the same format
Ease of manipulation	Extent to which data is easy to manipulate and apply to different tasks
Free-of-error	Extent to which data is correct and reliable
Interpretability	Extent to which data is in the appropriate languages, symbols, and units, and the definitions are clear
Objectivity	Extent to which data is unbiased, unprejudiced, and impartial
Relevancy	Extent to which data is applicable and helpful for the task at hand
Reputation	Extent to which data is highly regarded in terms of its source and content
Security	Extent to which access to data is restricted appropriately to maintain its security
Timeliness	Extent to which data is sufficiently up-to-date
Understandability	Extent to which data is easily comprehended
Value-added	Extent to which data is beneficial and provides advantages from its uses

1. The intrinsic category: It includes dimensions that express the natural quality of the data, such as accuracy, believability, objectivity, and reputation.

2. The contextual category: It expresses the fact that the quality of the data must be considered in a specific context. These dimensions include the amount of data, its completeness, relevancy, value-added, and timeliness.

3. The representational category: It refers to dimensions related to the format and meaning of the data, such as the ease of understanding, interpretability, and consistency.

4. The accessibility category: It refers to dimensions that express how data is accessible to users [Laranjeiro et al., 2015], including the ease of access and security.

Data quality in general and the DQD, in particular, have been widely studied in the literature [Omri et al., 2020]. These studies have attempted to discuss the most important DQDs and their interdependencies [Juddoo et al., 2018]. In [Islam, 2013], the authors proposed to discuss the timeliness dimension. They explored the impact of improving accuracy and completeness on the timeliness dimension. As a result, the authors proved that recently created data are more accurate. In the same context, Panahy et al. [Panahy et al., 2013] propose to study the dependencies between the four most important DQDs (according to the authors): accuracy, completeness, consistency, and timeliness. The obtained result proves a high dependence between these dimensions.

Despite this considerable number of studies in data quality and its dimensions, it still lacks the formalization of data quality [Cai and Zhu, 2015]. For example, the authors of [Cappiello et al., 2004] use the concepts of accuracy and completeness interchangeably. Moreover, accuracy has been used as precision and also semantic accuracy in [Caballero et al., 2014]. Dong et al. [Dong and Srivastava, 2015] explained the notion of accuracy as 'true value' in the context of data fusion, which refers to a process of integrating data

from various sources while maintaining a standard of data quality.

This lack of coherence in addressing the DQ and their dimensions is due to many studying contexts [Juddoo et al., 2018]. The DQ topic is addressed in different contexts such as the industrial domain [Omri et al., 2020], the web field [Batini et al., 2009] and the medical domain [Juddoo et al., 2018]. The following section introduces the data quality in the PHM concept.

2.5 Data quality in the PHM context

Data quality metrics differ from one application to another. Three data quality metrics are defined in [Jia et al., 2017] to characterize data for PHM applications. These metrics concern the aspects of detectability, diagnosability, and trend-ability [Jia et al., 2017]. Detectability refers to the fault detection task in the PHM framework, and it represents the ability of system abnormal behavior data to be detected and separated from the normal ones. Diagnosability fits the fault diagnosis within the PHM approach, and it means that data allow good separation of the various system failure modes. As for trend-ability, it concerns the degradation prediction, and it describes the ability of data to estimate accurate information about the remaining useful life (RUL) of the system [Lee et al., 2014b]. To measure these data qualities, Jia et al. [Jia et al., 2017] proposed using a statistical test based on the Maximum Mean Discrepancy (MMD) method used to evaluate the difference between two data distributions. Despite the obtained results, this method strongly depends on the used data modeling algorithm, ignoring the impact of the data quality. In coherence with this work, the authors in [Chen et al., 2013] define the cluster-ability as a metric that can be used in the PHM context. Cluster-ability measures data's predisposition to be clustered in natural groups without referring to the real data labels [Jain and Dubes, 1988].

Similarly, this method strongly depends on the used clustering algorithm. The proposed metrics describe the data quality in an aggregated way unrelated to the fundamental data quality issues (i.e., missing data, noisy data, incomplete data, etc.). Moreover, the authors in [Omri et al., 2019] propose a set of data quality requirements to be suitable with PHM applications. The data issues evoked in the mentioned article concern mainly: data volume, data accuracy, and completeness. Consequently, we propose in this work to classify the data problems according to these characteristics.

2.5.1 Data volume

One of the most critical factors that lead to a PHM project's failure is data unavailability. Data volume is the most critical data quality dimensions, and it concerns different aspects such as [Omri et al., 2020]:

- **The number of instances**: It refers to the existing volume of data (number of observations) that can be used to build a PHM model. This data quality is

measured by:

$$q_v = \mid R \mid \tag{2.1}$$

where $\mid . \mid$ refers the cardinality of the data space and R is the ensemble of objects that make up the database.

- **The imbalanced data**: It is a form of between-class imbalance that arises when one data class dominate over another class. It causes the machine learning model to be more biased towards majority class. The following metric is used to quantify this aspect of data problem:

$$q_{Im} = 1 - \frac{\mid o \in S \mid}{\mid R \mid} \tag{2.2}$$

where o is an observation and S is the objects ensemble of the subsampled class.

2.5.2 Data accuracy

Data accuracy is one of the most frequently cited dimensions of DQ in the literature. In [Batini and Scannapieco, 2016], authors define accuracy as the "distance" between the data or information and the world reality they describe. Data accuracy could control:

- **The outlier data:** It is one of the well known data quality problem that refers to data objects that do not correspond to expected behaviors [Hodge and Austin, 2004]. It is defined by:

$$q_o = \frac{\mid o \in A \mid}{\mid R \mid} \tag{2.3}$$

where A is the ensemble of outlier observations in the dataset.

- **The noisy data:** It concerns data that are recorded with an error compared to the world reality they describe. From a logical point of view, a value is considered noisy only if it impacts the detection result. We propose here to quantify the accuracy ratio as follows:

$$q_n = \frac{\mid o \in N \mid}{\mid R \mid} \tag{2.4}$$

where N is the ensemble of noisy observations for a specific feature X_i,$i = 1, \ldots, n$.

2.5.3 Data completeness

Completeness is the data dimension that deals with the problem of missing data. We here differ between two types of missing data [Omri et al., 2021]:

- **Partially missing data:** It evaluates the ratio of missing values for a variable. Thus, completeness is explained in this case as the percentage of available values for a variable. The completeness ratio is calculated using the following metric:

$$q_m = \frac{\mid o \in M \mid}{\mid R \mid} \tag{2.5}$$

where M is the ensemble of missing observations for a specific feature X_i,$i = 1, \ldots, n$.

- **Completely missing data (Insufficient features):** This case is discussed in [Omri et al., 2020], and it concerns the case where one or more features are completely missing due to the absence of sensors or the fact that they are not measurable. Insufficient features ratio is quantified using:

$$q_{ins} = \frac{n}{d} \tag{2.6}$$

where n is the number of saved features X_i,$i = 1, \ldots, n$ and d refers to the number of identified variables during the data inventory step and that describe the system Σ.

In the PHM context, data quality is an essential topic that should be considered to improve the obtained results [Omri et al., 2020]. In this context and before investing in advanced sensing technologies, it is more efficient to apply algorithms to deal with data quality issues. The following section presents a brief review of data quality improvement techniques that deal with the previously studied data quality issues.

2.6 Review of data quality improvement techniques

In this section, the techniques to deal with the previously detailed PHM data quality issues are reviewed.

2.6.1 Imbalanced data

Imbalanced data is one of the most critical data problems for data analysis task. For that purpose, many techniques to deal with this issue have been proposed in the literature. These techniques can be divided into four main categories [Galar et al., 2011]:

- *Data level* techniques aim to pre-process the data and make the classes more balanced, allowing the use of standard data analysis algorithms.
- *Algorithmic level* methods consist of developing sophisticated classification algorithms to deal with this imbalanced data problem.

- *Cost − sensitive* methods belong to the first two categories. They involve data modification by adding costs to instances, introducing high costs of misclassification for minority classes, and modifying the training algorithm accordingly.

- *Classifier ensembles*, inspired by human reasoning, this strategy combines the results of several imbalanced data techniques to build a new classifier that surpasses them.

Table 2.2 presents a brief review of the most used techniques.

TABLE 2.2: Summary of imbalanced data processing techniques.

Category	Algorithm	Remarks
Data level	SMOTE [Chawla et al., 2002]	- The minority class is over-sampled - New samples are generated and not replicated - Generated samples are introduced along the line segments joining k-nearest neighbors of the minority class samples
	CBO [Zheng et al., 2016]	- K-means technique is used to cluster data from each class separately - Over-sampling is performed in each cluster
	GAN [Mariani et al., 2018]	- Deep learning generative models are modelled as neural networks that take as input a simple random variable and that return a random variable that follows the targeted distribution
Algorithmic level	SCOP-SVM [Cui and Xia, 2017]	- Second-order cone programming support vector machines - Linear programming is used to improve the accuracy and the robustness of SVM - This algorithm is only usable for imbalanced data
	NBSVM [Datta and Das, 2015]	- Near-Bayesian Support Vector Machine (NBSVM) - Reduce misclassification cost due to rare class
	SBHD [Beyan and Fisher, 2015]	- Similarity-Based Hierarchical Decomposition (SBHD) method - Works effectively when data is highly overlapping, and classes are more imbalance
Costsensitive (CS)	CS Ad-Boost [Sun et al., 2007]	- These kind of algorithms is based on the Adaptive Boosting technique - A cost-sensitive boosted ensembles performed better than plain boosting - AdaC1, AdaC2, AdaC3 and AdaCost are CS boosting algorithms
	CS ANN [Tsai et al., 2009]	- Cost sensitivity can be introduced to neural networks in the probabilistic estimate, the neural network outputs can be made cost-sensitive, cost-sensitive modifications can be applied to the learning rate and the error minimization function can be adapted to account for expected costs
	CS DT [Krawczyk et al., 2014]	- Cost-sensitive fitting can take three forms: adjustment of the decision threshold, cost sensitive split criteria at each node and the application of a cost-sensitive pruning to the tree
Classifier ensembles	Bagging [Hido et al., 2009]	- Generate 'n' different bootstrap training samples with replacement - Train the used algorithm on each bootstrapped data separately - Obtained performance are highly depend on the used classifiers - Results are presented as an aggregation of the different predictions
	Boosting [Galar et al., 2011]	- Combine weak classifiers to build a more accurate classifier - Algorithms are trained sequentially in a adaptative way - Results are obtained by combining predictions following a fixed strategy - Ada-Boost, Gradient-Boosting and XG-Boost are the most used boosting techniques

In this work, we focus only on the data-level methods. Three algorithms from Table 2.2 were considered relevant and used in this work. SMOTE and CBO were selected because they are widely used in the literature, and their performance has been proven. In contrast, the GAN technique was selected because it is a recent method that has demonstrated satisfactory performance in dealing with this data problem. For that, a brief description of these techniques is given next.

Synthetic Minority Over-Sampling Technique (SMOTE) [Chawla et al., 2002]
This method is one of the most used oversampling techniques in the literature. It creates new observations along the lines of a randomly chosen point and its k-nearest neighbors. However, the generated similarities between instances call into question the performance of this technique.

Clustering-Based Oversampling (CBO) [Zheng et al., 2016] To avoid creating similarities in oversampled classes, the CBO technique applies the K-means clustering algorithm independently to identify the dataset clusters. These classes are then oversampled to balance the data. In this second phase, traditional oversampling algorithms can be used, such as SMOTE or random oversampling techniques. Despite the use of the K-means method, it still possible to over-fit the training data.

Generative Adversarial Network (GAN) [Mariani et al., 2018] This technique is decomposed from two neural networks: *Generator* and *Discriminator*. The first one allows to learn from the data and to generate new samples that follow the same distribution as the original samples. While the second one, the discriminator, assesses for each instance whether or not it is generated. The idea behind "adversary" is that a good generated results in a poor discriminator and vice-versa. For that purpose, the two neural networks must have similar skills.

2.6.2 Missing data

Missing values in the data can interfere with the execution of the data analysis task. To this end, many techniques for imputing missing data are available in the literature. Before detailing these techniques, it is a good idea to recall the types of missing data based on assumptions about the reasons for the missing data. Three main classes are identified according to the mechanisms of missing data [Little, 1988]:

- *Data missing completely at random:* Represents the fact that a certain missing value has nothing to do with its hypothetical value and with the values of other variables.
- *Randomly missing data:* Means that the probability of missing data depends on the observed data set, but is not related to the specific missing values expected to be obtained.
- *Non-random missing data:* If the characters in the data do not match these types, they fall into the category of non-random missing data.

The best way to deal with missing data is to work through the data collection process to avoid this problem. However, it is impossible to avoid the loss of some values. For this reason, several techniques for dealing with missing data have been proposed in the literature.

As shown in Table 2.3, techniques for dealing with missing data can be classified into three broad classes:

- *Conventional methods :* These techniques consist of leaving missing values or deleting them. The first family accepts missing data and generally involves the use of flexible data analysis algorithms that accept missing data. While the second family of techniques requires removing missing data, either totally or partially (only instances with a high percentage of missing values are removed).

TABLE 2.3: Taxonomy of missing data imputation techniques.

Category		Method	Remarks
Conventional methods	Deletion	Listwise [King et al., 1998]	- Delete all instances that contain missing values - Suitable only for low missing data fraction and high data volume - Information loss particularly in the in small data
		Pairwise [Shi et al., 2020]	- Delete only the instances with high levels of missing data - Suitable only for low missing data fraction and high data volume
	Ignoring	[Nakagawa and Freckleton, 2008]	- Do nothing
Imputation methods	Statistical methods	Mean, Median and Mode imp. [Donders et al., 2006]	- Replace missing values with the mean or median for numerical variables and the mode for categorical variables - Change the characteristics (mean, variance) of the original variables
		Hot Deck [Myers, 2011]	- Missing values are imputed from randomly selected records
		Cold Deck [Chhabra et al., 2017]	- Missing values are imputed from similar records
		MICE [Van Buuren and Oudshoorn, 1999]	- Simple imputation models are used to make multiple imputation for each missing value - Analyze the generated completed data sets - Integrate the analysis results into a final result - Operates under the assumption that data are missed at random
		Regression imp. [Arteaga and Ferrer, 2005]	- Regression model is used to predict and replace missing data
	Machine learning methods	Decision Tree [Rahman and Islam, 2011]	- The DT algorithm is used to predict and replace missing data
		SVM imp. [Mallinson and Gammerman, 2003]	- The SVM algorithm is used to predict and replace missing data
		KNN imp. [Malarvizhi and Thanamani, 2012]	- The KNN algorithm is used to predict and replace missing data - Compatible with quantitative and qualitative data - Sensitive to outlier data - Computationally expensive
		ANN and DL [Beaulieu-Jones et al., 2017]	- The DL algorithms are used to predict and replace missing data
Likelihood-based methods	EM [Lin, 2010]		- Operates under MAR and MCAR assumptions
	FIML [Graham, 2003]		- Operates under MAR and MCAR assumptions

- *Imputation methods :* This family of techniques involves replacing missing data with synthetic values. These values are generated according to a procedure based on statistical methods or machine learning techniques.

- *Likelihood-based methods:* In these techniques, the assumption that the observed data is a sample drawn from a multivariate normal distribution is relatively easy to understand. Once the parameters are estimated using the available data, missing data are estimated based on the newly estimated parameters.

2.6.3 Noisy data detection

The existence of noise in the data can strongly affect the performance of data analysis algorithms. To this end, many techniques have been proposed to solve this problem. As shown in Figure 2.4, outlier detection methods can be divided into two main groups: (*i*) *detection Supervised* and (*ii*) *detection unsupervised* [Suri et al., 2019]. The first requires labeled data to train a model that classifies each instance as an outlier or normal instance. While outliers are generally rare, supervised detection algorithms face a data imbalance problem that calls into question their effectiveness in dealing with the outlier detection problem. On the other hand, unsupervised techniques detect outliers as a deviation from normal data based on a predefined deviation metric. Despite the high rate of false positives, unsupervised techniques remain the most widely used in the literature [Campos et al., 2016]. For this purpose, only unsupervised outlier detection methods are considered in this study.

Unsupervised techniques are based on an appropriate deviation from the normal data metric to declare an object as an aberrant or normal object. These methods can be parametric or non-parametric (see Figure 2.4). Parametric techniques are generally based on statistical distributions for data modeling. An object is declared an outlier or not with respect to its position in the data model's probability regions. In the same

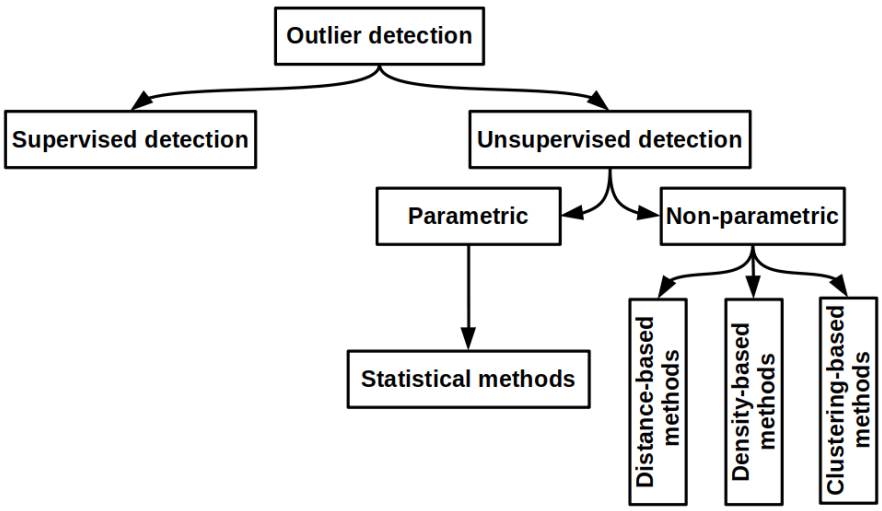


FIGURE 2.4: Outlier detection techniques cartography.

context, non-parametric methods are divided into three main categories: distance-based, density-based, and classification-based methods [Campos et al., 2016].

Distance-based techniques involve consulting the neighborhood of objects, and an object is declared as an outlier if there are not enough objects in its neighborhood. Thus, these techniques require a radius r and a threshold p $(0 < p < 1)$ to characterize the neighborhood and the necessary objects, respectively. Mathematically, an object o is identified as aberrant if the equation 2.7 is satisfied.

$$\frac{|o' \in D \mid dist(o,o') \leq r|}{|D|} \leq p \tag{2.7}$$

where D is the set of objects o' that make up the database and $dist$ $(.\ ,\ .)$ is a measure of distance.

Density based techniques are based on the assumption that the density around a normal object is similar to that around its neighbors while the density of an aberrant object is different. To do this, we define $dist_k(o)$ the distance between an object o and its k^{th} neighbor. Thus, the neighborhood of o is defined as $N_k(o) = [\, o' \in D \mid dist(o,o') \leq dist_k(o)\,]$. Thus, an instance is declared as an outlier if the equation 2.8 is satisfied.

$$\frac{1}{|N_k(o)|} \times \sum_{o' \in N_k(o)} dist(o,o') > P \tag{2.8}$$

where P is a predefined threshold and $dist$ $(.\ ,\ .)$ is a measure of distance.

TABLE 2.4: Summary of outlier data detection techniques.

Group	Category	Algorithm	Reference	Data type	Data volume	Data dimension
Supervised detection		APD	[Das et al., 2008]	Categorial	Large	Multivariate
		RF	[Liaw et al., 2002]	Both	Large	Multivariate
		SVM	[Erfani et al., 2016]	Both	Low/moyenne	Multivariate
Unsupervised detection	Statistical	GMM	[Dang et al., 2015]	Numeric	Low	Multivariate
		HBOS	[Tang et al., 2015]	Both	Large	Multivariate
		RLOD	[Goldstein and Dengel, 2012]	Numeric	Large	Uni-variate data
	Distance-based	kNN	[Du et al., 2015]	Both	Large	high
		Abstract-C	[Yang et al., 2009]	Both	Large	Multivariate
		COD	[Bifet et al., 2010]	Both	Large	Low
	Density-based	LOF	[Breunig et al., 2000]	Numeric	Low/moyenne	Multivariate
		RDOS	[Bai et al., 2016]	Both	Large	Multivariate
		INFLO	[Jin et al., 2006]	Both	Low	Multivariate (d12)
	Clustering-based	CBLOF	[He et al., 2003]	Both	Large	Multivariate
		Isolation forest	[Das et al., 2016]	Both	Large	high
		K-means	[MacQueen et al., 1967]	Both	Large	Multivariate

Clustering-based techniques are motivated by the fact that there are fewer aberrant objects than normal objects. Thus, objects in small clusters are identified as outliers.

Many algorithms have been proposed in the literature to deal with the problem of detecting unsupervised outliers. A review of these techniques is presented in the Table 2.4 by grouping them according to the previously proposed classification.

2.7 State of the art synthesis

Based on the previously conducted data quality literature review, there exist two solutions to overcome data quality issues such as software and technological data quality improvement. As mentioned above, the first one is about data quality improvement algorithms, while the second refers to the installation of more sophisticated sensors.

Recall that this work concerns SMEs, which are generally limited in terms of financial resources. Therefore, technological solutions to improve existing data quality or to collect new variables are not considered because they can generate high costs. Therefore, technological solutions are not considered in this thesis.

Concerning the software improvement techniques, they are applied in the various data preprocessing tasks in this work. However, they are not applicable when variables are not recorded, or they are highly damaged [Omri et al., 2020] which is the case in SMEs where data acquisition infrastructure is not well developed. [Omri et al., 2021]. **Due to these constraints, traditional data quality improvement solutions are not suitable for SMEs.**

As shown in Figure 2.5, the main research question appears in the data quality management context:

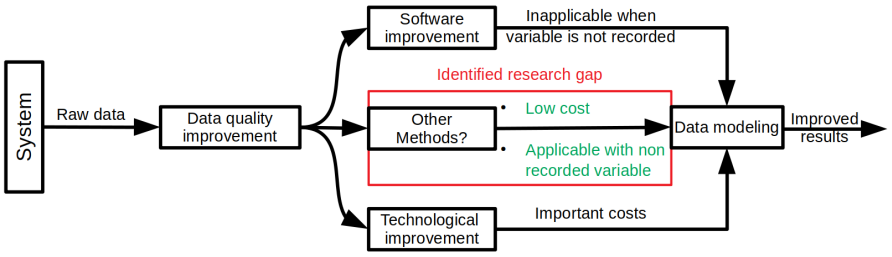


FIGURE 2.5: Research gaps identification on the data quality improvement literature.

- **How to improve the quality of highly damaged variables at a low cost?**

The next chapter answers this research question by proposing a new low-cost data improvement technique.

2.8 Conclusion

Recall that this work's objective is to implement Industry 4.0 in SME using an adapted PHM approach. To do, the industrial data digitization problem is addressed in this chapter. In particular, the data quality problem is presented, and its impact on the PHM process is discussed. Besides, the data quality improvement techniques are reviewed, and their performance is discussed. The studied data problems are identified according to a real analysis conducted in the SCODER company.

Based on this real analysis, it is proved that industrial data digitization is a complex subject that is about data quality and quantity. Indeed, digitized data availability is a problem that concerns most companies and, in particular, SMEs where the data acquisition infrastructure is not well developed. Thus, data on the production process are not recorded, which impacts the PHM results. As a consequence, traditional data improvement techniques are not applicable in the case of SMEs. On the other hand, SMEs have a large amount of data in the form of know-how that can be used to overcome this problem. Thus, it is necessary to integrate staff knowledge to improve data quality, the PHM results and to facilitate the Industry 4.0 implementation. The following chapter presents a new knowledge-based approach for data quality improvement.

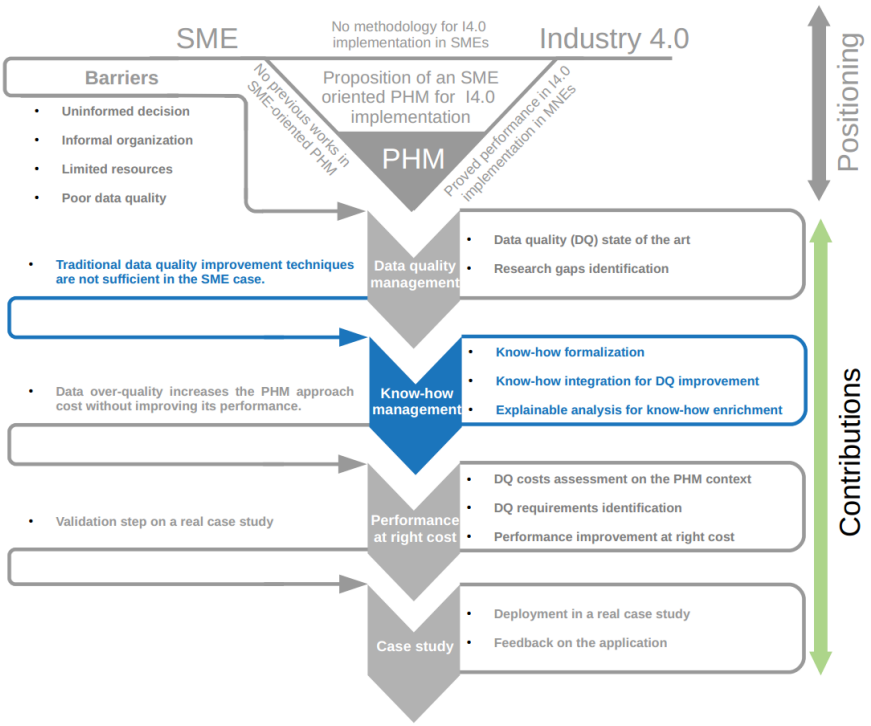
Chapter 3

Data quality management based on Knowledge oriented methodology

"The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge."

-Stephen Hawking

Graphical abstract.



Contents

Contributions

This chapter introduces a generic formalization of the operators' know-how to use it to improve data quality. For this purpose, human knowledge types are defined, and their integration modes in the data management process are discussed. This study is integrated into a new closed-loop (data quality - knowledge) for data quality improvement and human knowledge enrichment. The proposed approach is applied in the SCODER case study to validate it and assess its applicability.

3.1 Introduction

Industrial data digitization is a complex subject that concerns data quality and quantity. This problem is intensified in SMEs' case, where the data acquisition infrastructure is not so developed. As a result, part of the data on the production process is not recorded. These limitations result in a reduction in SMEs' ability to adopt new technologies, which reveals industry 4.0. However, SMEs have a massive amount of data in the form of remarkable know-how that is not efficiently valorized. Indeed, SMEs' advantage is that operators are close to the production process, which allows them to identify their needs and prioritize them in terms of urgency. Besides, the staff's versatility allows them to analyze each problem regarding its impact on the entire production process. As a result, they can extract the necessary and relevant data to steer digitization projects and ensure knowledge capitalization. This chapter introduces a generic formalization of the

operators' know-how to improve data quality and improve the analysis results. For this purpose, human knowledge types are defined, and their integration modes in the data management process are discussed. This study is valorized in a new knowledge-based data quality improvement framework. In a second step and to ensure efficient knowledge management, the obtained analysis results are explained and used to enrich the human knowledge base. This same knowledge is applied to improve data quality in future analysis tasks.

In summary, this work aims to develop a knowledge-based methodology for efficient data quality management. The decisions resulting from the analysis of these data are returned with a concern for explainability to enrich the human knowledge base used again to improve data quality in future analysis tasks. The rest of this chapter is organized as follows. The problem statement and the proposed solutions are drawn in section 3.2. Sections 3.3 and 3.4 present state of the art on the formalization of human knowledge and the explainable machine learning. Section 3.5 details the proposed approach. This framework is applied to a real case study, the results of which are presented in section 3.6. The conclusions and perspectives of this work are detailed in section 3.7.

3.2 Problem statement and proposed approach

As noted in the previous chapter, Industry 4.0 is a data-based revolution where data quality management is one of the most critical challenges. This section presents the data quality problem in SMEs while proposing a new approach to overcome it.

SMEs' data acquisition infrastructure is not well developed, which affects the collected data quality. As a result, highly damaged variables are very common in SMEs' data. As detailed in the previous chapter, the main research question in this domain is how to improve highly damaged data at a low cost. In this chapter, we propose improving the data quality using the operators' know-how to improve the obtained results. These results are explained to enrich the human knowledge base and thus ensure efficient knowledge management. This same knowledge is applied to improve data quality in future analysis tasks. However, two main problems arise in this context:

- How to integrate the know-how of the staff?
- How to enhance an efficient knowledge capitalization?

We face here two complex problems in the data analysis domain. As they are introduced in the literature, these problems are the informed [von Rueden et al., 2019] and the explainable [Guidotti et al., 2018] machine learning tools. The explanation term refers to the used machine learning algorithm's ability to explain the obtained results and models. This makes the algorithm more transparent and therefore builds trust between it and its users. While "informed machine learning" refers to integrating human knowledge in the learning process to improve it.

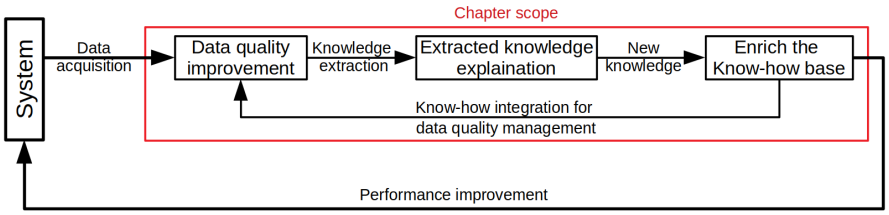


FIGURE 3.1: The proposed knowledge-based data quality improvement approach.

As shown in Figure 3.1, the proposed approach consists of four main steps: (i) improve data quality using human knowledge, (ii) extract knowledge from these data, (iii) explain the extracted knowledge to enrich the human knowledge base, and (iv) use this knowledge to improve data quality in future data analysis tasks. To sum up, this chapter proposes a closed-loop (data quality - knowledge) for data quality improvement and human knowledge enrichment and capitalization. To do this, we propose in the following sections to detail the state of the art of techniques used to explain the results of the data analysis. Besides, the types of human knowledge are presented as well as the methods of their integration into the data analysis process.

3.3 Human knowledge formalization: related works

As shown above, human knowledge is an important source of information that should be used to guide the used machine learning tools. This information enables efficient learning from the data and ensures more accurate results. In this context, many approaches turn human knowledge into a useful source of learning for the used machine learning algorithms [von Rueden et al., 2019].

This section proposes a literature review of these approaches while detailing the different kinds of human knowledge and formalizing them for informed learning.

3.3.1 Human knowledge types

Humans are the most intelligent creatures on earth where their knowledge remains complex and unknown in terms of quantity. Many studies have been conducted in the literature to characterize this knowledge from different perspectives (e.g., sociology, psychology points of view, etc.). These studies propose several ways to categorize human knowledge. In line with these categorizations, we propose here to classify human knowledge in three main types: (*i*) General knowledge, (*ii*) Scientific laws, and (*iii*) Expertise.

General knowledge [von Rueden et al., 2019]

This type of human knowledge refers to basic information known and accessible to all humankind. This knowledge is developed by human intuition and passed on from one generation to another; for example, the temperature will rise in summer.

Scientific laws [Srinivasan et al., 2020]

As the title indicates, this knowledge represents the set of facts and laws explicitly validated by experiments. Usually, these laws are represented in the form of mathematical equations that relate facts to consequences. Many areas are described by these universal laws, such as physics, chemistry, etc.

Expertise [Spinner et al., 2019]

The latter types of human knowledge reflect the particular know-how developed by a person in a particular field. Usually, this wisdom relates to the field of work in which a person spends much time. Observations are the main source of this knowledge and can be shared among teammates orally or as documents [Omri et al., 2019]. In addition to know-how, the category of expertise may include information accessible to human senses, such as heating a machine or a sound noise. Unlike scientific laws, this knowledge (if documented) is saved in the form of texts.

3.3.2 Formalization of the human knowledge

As mentioned in the previous paragraph, human knowledge is saved in the form of text, and in the best of cases, it is represented by mathematical equations. However, it still difficult to communicate human wisdom to a machine. Natural communication between humans is not understandable by machines. For that purpose, many approaches to formalizing human knowledge in machine-understandable forms are developed in the literature. We propose here to classify these approaches in three main categories: (*i*) Mathematical models, (*ii*) Logic rules, and (*iii*) Statistical relations.

(i) Mathematical models

Mathematical models formalize knowledge as functions or as differential equations. Functions relate a set of variables to output, or they can be used as constraints to reduce the solution's space. In this context, we can cite the Ohm's law ($V = R \times I$) which relates the voltage V to the current I and the resistance R. However, differential equations describe the relationship between functions, and they allow the formalization of human knowledge as a compact mathematical model. We can also cite Newton's second law, which is formalized as a differential equation. From the writer's point of view, mathematical models are the most suitable to communicate between humans and machines since they are understandable by the two sides. However, not all human knowledge can be modeled mathematically. Thus, other formalization methodologies are developed.

(ii) Logic rules

Over time, human knowledge is saved as texts. However, natural language phrases are not understandable by machines. Thus, logic is used to transform facts and consequences represented by sentences to formal logic rules [von Rueden et al., 2019].

(iii) Statistical relations

As mathematical functions, statistical relations relate data variables to each other. The variables can be correlated, or they can be defined as a statistical distribution. Moreover, statistical relations can refer to conditional independence or correlation structure of random variables or even a complete description of the joint probability distributions [von Rueden et al., 2019].

3.3.3 Informed learning approaches

We now come to describe the different methodologies of informed learning. As mentioned above, informed learning is defined as the combination of data and human knowledge for a more efficient learning process. This kind of learning get more attention in recent years, and several techniques have been proposed in the literature. Before detailing these techniques, we first describe the learning process and then relate each method to a step in that process. Globally, a learning process can be divided into three main steps: (A) Data preprocessing, (B) Data mining, and (C) Results. Thus, human knowledge can be integrated into the learning process throughout these stages. We propose here to classify informed learning approaches according to the impacted stages.

(A) Informed data preprocessing:

Data preprocessing is an important task in the learning process, and it consists of preparing data for the data mining phase. This task is divided into many steps, such as data cleaning, data transformation, and feature selection. Data cleaning and transformation are methods used to remove outliers and standardize the data so that they take a form that can be easily used to create a model. While features selection consists of removing the non-pertinent variables to reduce the data dimensions and then accelerate the learning process and improving the result's accuracy.

In this context, human knowledge can be used in these different tasks. Indeed, prior knowledge of the studied problem can lead to the features selection phase. Also, it can assess the accuracy of the collected data and improve it accordingly. Additionally, it can be used as a second data source by using a mathematical function to generate new variables from existing ones or generate new data using statistical distributions [Ladický et al., 2015]. Last but not least, a human can feed a learning algorithm through a collection of rare observations and events that are difficult to formalize in the input data.

(B) Informed data mining:

Data mining consists of learning from data and extracting knowledge from them. To do this, artificial intelligence algorithms are used to predict outputs from a set of input variables. As an example of algorithms that can be used, we cite artificial neural networks

(ANN), support vector machine (SVM), decision trees (DT), etc.

At the level of the data mining, the human knowledge can identify the parameters set of the used algorithm [Zemouri et al., 2019] such as the architecture of an ANN or the deep of the trees of DT, etc. Moreover, prior knowledge of the studied problem can modify the loss function according to the final objective [Krawczyk et al., 2014].

(C) Informed results:

Results are the objective of the learning process. They represent a set of rules linking inputs to outputs. These rules can be implicit or explicit, simple or complex, and accept a certain uncertainty order.

At this level, humans can intervene to assess the consistency of the extracted rules by comparing them to existing scientific laws or by merely using their expertise. Additionally, known rules that are not learned from the data can be injected into the final results to document them.

As displayed in Table 3.1, human knowledge can be integrated into the learning process throughout its different steps. However, some tasks are automated by more sophisticated algorithms to facilitate the user's mission. Growing algorithms are used to adapt the ANN architecture to the problem complexity [Zemouri et al., 2019]. For the feature selection step, many techniques to automatically accomplish this mission are developed and applied in many domains [Zemouri et al., 2018]. Other methods are also developed to learn efficiently from imbalanced data [Tsai et al., 2009] or to automatically assess the results consistency [Mariani et al., 2018]. For that purpose, we propose here to integrate human knowledge in the generation of new data and improve their quality. However, human knowledge is limited in some domains and may be biased. For that, this knowledge must be sustainably enriched and evaluated. This work proposes to explain the data analysis results to enrich the human knowledge base with new reliable information. The following section presents a brief review of existing explainable data analysis techniques.

3.4 Overview of explainable data analysis techniques

Many precise decision support systems have been built like black boxes in recent years, hiding their internal logic from the user. This lack of explanation is both a practical and an ethical problem. The literature reports many approaches to overcome this critical weakness. This section proposes to study this problem while specifying the needs in terms of explanations and the existing techniques.

TABLE 3.1: Human Knowledge Integration examples in informed learning process.

Human knowledge type	Formalization approaches	Examples	Knowledge integration examples		
			Data preprocessing	Data mining	Results
General knowledge	Mathematical models	Engine $\in Cars$	-	-	Check the engine in case of problems
	Logic rules	if (Summer) then: Temperature rise	Control temperature in the summer	-	Ventilate machines in summer
	Statistical relations	Data are imbalanced	Use data balancing techniques	Adapt the loss function	-
Scientific laws	Mathematical models	Ohm's Law: V = R x I	- Generate V from R and I variables .- Assess the accuracy of I variable using the formula I=V/R.	-	V = R x I
	Logic rules	if (T>100 °C) then: water vaporize	-	-	if (T>100 °C) then: water vaporize
	Statistical relations	P(A inter B) = 0	-	-	Events A and B can't be simultaneous
Expertise	Mathematical models	The problem is very complex	-	Use deep architecture of ANN	-
	Logic rules	if (T>80 °C) then: machine failure	-	-	Stop the machine when: T>80 °C
	Statistical relations	Correlation (Result, var2) = 1	Feature selection: var2 is very important	Use linear regression algorithm	Assess the quality of the prediction

3.4.1 Explanation needs

The explanation is not necessary either because (1) there are no significant consequences for unacceptable results or (2) the problem is sufficiently well studied and validated in real applications to be confident in the decision system, even if the system is not perfect. The need for interpretability arises from an incomplete formalization of the problem, creating a fundamental obstacle to optimization and evaluation. Note that incompleteness is distinct from uncertainty: the fusion estimate of a missile position can be uncertain, but this uncertainty can be rigorously quantified and formally reasoned about it. In machine learning terms, we distinguish between cases where the unknowns lead to quantified variance, for example, when trying to learn from a small amount of data or with a lack of completeness which produces some unquantified bias. Below are some illustrative scenarios:

- **Scientific understanding:** The goal of a machine learning process is to extract knowledge. Thus, **explanation can be helpful to enrich the human knowledge base**.
- **Security:** For complex tasks, the end-to-end system is rarely completely testable; one cannot create a complete list of scenarios in which the system may fail. Listing all possible exits, including all possible entries, is computationally or logistically impractical, and we may not be able to report all unwanted exits.
- **Ethics:** Human beings may want to guard against certain types of discrimination. For example, one may want a "fair" classifier for loan approval. Even though we can encode protections into the system for specific protected classes, there may be biases that we have not taken into account a priori.

In this thesis work, explainability is used to enrich human knowledge and improve our mastering level of the problem. The following paragraph details the most used explainable techniques.

3.4.2 Explanation models

Increasingly, the explanation of black-box decision systems has attracted more attention. This need for explanation is generally due to incomplete problem formalization, creating a fundamental obstacle to optimization and evaluation. Thus, many techniques have recently been proposed to explain black-box decision systems [Guidotti et al., 2018]. In this context, the authors of [Tan et al., 2020] applied decision trees to explain neural network decisions. Indeed, classification rules have been widely adopted to explain the decisions of neural networks [Johansson et al., 2004] and SVMs [Fung et al., 2005]. These techniques are used to generate a **global explanation** of the used black-box model. When the training dataset is available, they can be used as completely transparent classifiers.

Other approaches tackle explaining the **local behavior** of a black-box model [Guidotti et al., 2018]. In other words, they explain the decision assigned to a specific data observation. There are two types of approaches: model-dependent approaches and agnostic approaches. In the first category, most of the articles aim to explain neural networks. They base their explanation on salience masks, that is, a subset of the instances that explains what is primarily responsible for the prediction [Zhou et al., 2016]. Examples of salience masks are parts of an image or words or phrases in the text. On the other hand, agnostic approaches provide explanations for any black-box model. In [Ribeiro et al., 2016], the authors present the LIME techniques, which starts from instances generated randomly in the vicinity of the instance to be explained. The method deduces linear models from them as well as understandable local predictor models. The feature's importance in the linear model represents the explanation ultimately given to the user. As a limit of this approach, the randomly generated data instances do not fit the problem reality. Therefore, the linear classifiers derived from them may not correctly characterize the black-box model's reasoning principle.

LIME extensions using decision rules (called Anchors) and expression trees are presented in [Ribeiro et al., 2018] and [Singh et al., 2016] respectively. The Anchors extension [Ribeiro et al., 2018] uses a specific algorithm that randomly constructs the anchors with the highest coverage and respecting a precision threshold. In [Singh et al., 2016], the authors take a simulated annealing approach that randomly increases, decreases, or replaces nodes in an expression tree. The adopted neighborhood generation process is the same as that of the LIME technique. Another crucial weak point of these approaches is the need for user-specified parameters for the desired explanations: the number of features, the level of precision, the maximum depth of the expression tree [Ribeiro et al., 2018].

Despite the multitude of explanation models, their explanation quality remains insufficient. They are divided into accurate local explanations and inaccurate global explanations. Thus, the objective is to propose a new approach that can provide precise global explanations. However, we should be able, firstly, to assess the quality of an explanation. In this context, three main assessment approaches are proposed: application-based [Antunes et al., 2008], human-based [Lakkaraju et al., 2016], and function-based [Freitas, 2014] . We here consider that this research domain is out of the scope of this thesis work. Thus, we propose to assess the quality of an explanation directly via the industrial application.

Next in this chapter, the informed and explainable learning concepts are introduced in the proposed knowledge-based DQ improvement. The decisions resulting from this approach are explained to ensure human knowledge enrichment. The following section details the proposed framework.

3.5 A proposed data quality management methodology

This section details the different steps of the proposed knowledge-based approach for data quality management.

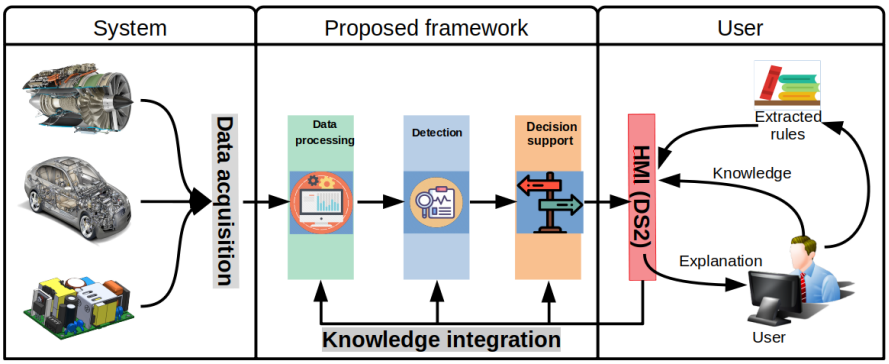


FIGURE 3.2: The proposed knowledge-based approach for data quality management.

As shown in Figure 3.2, the proposed approach consists of an interactive framework that allows communication between the user and the black-box machine learning algorithm. This interaction is represented in (i) integrating user expertise for DQ improvement and (ii) explaining the analysis results for knowledge enrichment. Thus, human knowledge can be validated through this process to confirm (or decline) it. The capitalized knowledge is then stored in a library of rules that will be used in the future to support decision-making and explain strange phenomena.

3.5.1 Knowledge integration for data quality management

This section presents the phases of knowledge integration for data quality management. As shown in the table 3.1, human knowledge can be integrated into the learning process throughout its various stages. However, some tasks are automated by sophisticated algorithms to facilitate the user's mission. Scalable algorithms are used to adapt the ANN architecture to the problem complexity [Zemouri et al., 2019]. For the variables selection step, many variables selection techniques allowing to accomplish this mission automatically are developed and applied in many fields [Zemouri et al., 2018]. Other methods are being developed to effectively learn from imbalanced data [He and Garcia, 2009] or automatically assess the results consistency. For this, we propose, here, to integrate human knowledge in data quality improvement. As detailed in the previous sections, there are several ways to integrate human knowledge into the learning process. Here, we propose to valorize human knowledge in the data quality improvement according to four main ways:

- **Feature importance definition:** Continuing with the previous element, the user understands the problem, its causes, and its consequences. He generally links the

problem to a set of variables that can affect it (or the variable with no influence). Thus, this knowledge will be introduced into the data analysis phase to improve its results.

- **New variables definition:** Thanks to his expertise, the user can judge that the fusion of a set of variables can generate a new variable more relevant for the description of the problem studied. Thus, the user is responsible for defining the calculation function of such a variable.
- **Variables description:** It refers to statistical description allowing information to be obtained on each variable separately. We distinguish two types of variables: (i) quantitative variables (continuous values) and (ii) qualitative variables (the number of possible values is limited). A quantitative variable's description is based on the following statistics: mean, median, variance, standard deviation, quantiles, min, and max. The description of a qualitative variable is much more summary. Once the variable modalities have been identified, it is a matter of identifying the mode and studying the proportions associated with each modality. Of course, this information can be calculated automatically. However, when we do not have a history of recorded data (which is the case of SMEs), the user's contribution becomes essential to have a complete idea about each variable.
- **Statistical relationships definition:** Statistical relationships link variables to each other. The variables can be correlated, or they can be defined as a statistical distribution. Also, statistical relationships can refer to conditional independence or even a complete description of joint probability distributions. Like the variables descriptions, this information can be calculated automatically. However, we are interested in cases where there is no history of recorded data.

On the other hand, there remains another challenge which is the facilitation of this task for a non-specialized user. To do this, we propose to formalize this work in an ergonomic graphical interface where the user can choose between several options to communicate his knowledge. This interface allows the user to choose between mathematical models, statistical relations, or logical rules to integrate his knowledge. Then, he indicates the variables concerned by this model. Finally, he defines the model through a set of parameters. In this context, it should be noted that this knowledge can contain biases (even false information) hence the need to evaluate it in a collaborative approach with the user. This layer aims to enrich our knowledge with reliable information.

As detailed above, user knowledge is used to improve data quality and improve the analysis phase results. In this work, the data analysis phase is not considered since many works deal with it. We propose applying traditional techniques such as ANN, DT, and SVM for the data analysis phase. The obtained results are explained to enrich the human knowledge base, which will be used in future data quality improvement tasks. The following subsection details the results explanation process.

3.5.2 Results explanation for know-how enrichment

This paragraph proposes to detail the explanation process. As shown in Section 3.3, several explanation techniques are proposed in the literature. These techniques are divided into accurate local explanations and inaccurate global explanations. In this work, we propose a semi-global explanation algorithm that takes advantage of these two techniques. For an explanation task, the most important thing is to determine a clear and understandable rule allowing to separate the different classes. As shown in Figure 3.3, the proposed approach is based on the generation of new instances at the boundary between classes. These new instances are then used to separate the classes linearly.

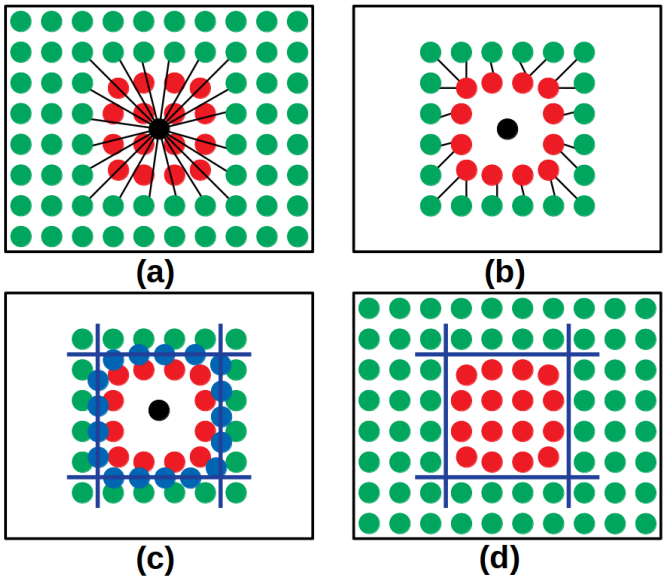


FIGURE 3.3: Explanation process details.

Let $\zeta(.,.)$ be a function which takes as inputs a black box model b and an observation of the system X_i for $i = 1, \ldots, l$. The function $\zeta(.,.)$ is called an explanatory model when it relates the decision $b(X_i)$ to a physical reality of the system. The proposed approach can be summed up in four steps:

- Identification of the decision plan.
- Data generation.
- Generated data authenticity assessment.
- Decision explanation.

These four steps are detailed in the next paragraphs.

Step 1: Identification of the decision plan

A decision plan is a plan separating the two classes closest to the instance X_i whose decision we want to explain. For a data point X_i whose decision we want to explain, the first step in identifying its decision plan consists of determining the k instances closest to it and which belong to the inverse class of that of X_i (see Figure 3.3 (a)). Using these k neighboring instances, the decision plan is determined. To do this, for each instance identified, we determine the instance closest to it and belonging to the same class as that of X_i (see Figure 3.3 (b)). For the rest of this work, the identified data points are called ***border points***.

Step 2: Data generation

As mentioned above, the second step in the explanation process is the new data generation.

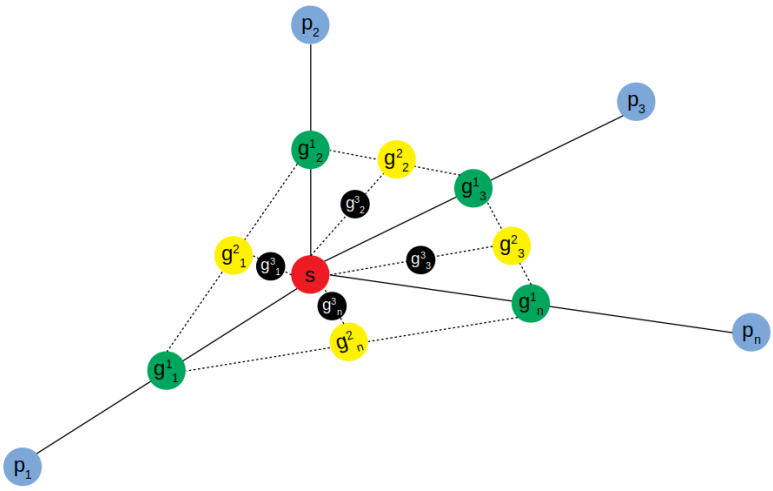


FIGURE 3.4: The data generation process. For a data point s (in red), the closest n points p_i (in blue) are used to generate a first data layer g_i^1 (in green). Then, this first layer is used to generate a second data layer g_i^2 (in yellow). Likewise, this second layer is used with the initial point s to generate the third layer g_i^3 (in black). This process is repeated until the generation of the required N data samples.

Figure 1 illustrates the generation process. In fact, the n nearest neighbors (p_i, $i = 1, \ldots, n$) of each data point s in the ***border points*** ensemble are selected to generate a first set of n new data samples g_i^1 as follows:

$$g_i^1 = \frac{p_i + s}{2}, \forall i = 1,..,n. \tag{3.1}$$

The generated data are then used to generate a second set of n new data samples g_i^2 as follows:

$$g_i^2 = \frac{g_i^1 + g_k^1}{2}, \forall k, i = 1,..,n. \tag{3.2}$$

where i and k are consecutive and $i \neq$ k .

Using the initial data point s and the points generated at the second level g_i^2, the third set of n new data samples g_i^3 are generated:

$$g_i^3 = \frac{g_i^2 + s}{2}, \forall i = 1,..,n. \tag{3.3}$$

The same process is repeated until the generation of the required N data samples. The Algorithm 1 presents the different data generation process stages.

Algorithm 1 The data generation process.

Initialization :
s = starting point;
n = number of neighbors;
N = number of required data samples;
G = generated data set;
j = 0: counter;

Algorithm:
while$(cardinal(G) < N)$ {
for (i= 1 .. n) {
$G = G + g_i^j$; }
$j = j + 1$; }
Required data are generated.

The generated data in this step will be used later in the results' explanations. Thus, it is important to check that the generated data are close to the real data to guarantee the provided explanations' quality. The following section presents the adopted approach to assess the authenticity of the generated data.

Step 3: Generated data authenticity assessment

After generating the neighbors of each instance, we focus here on validating the generated data authenticity. In fact, despite user intervention in improving the generated data quality, there are some deep features that are impossible to detect by a human. For this, a discriminator (see Figure 3.5) is trained to differentiate the original data from the fake ones.

The discriminator is used to evaluate the similarity between the new data and the real ones. This discriminator is inspired from the one used in the GAN neural network

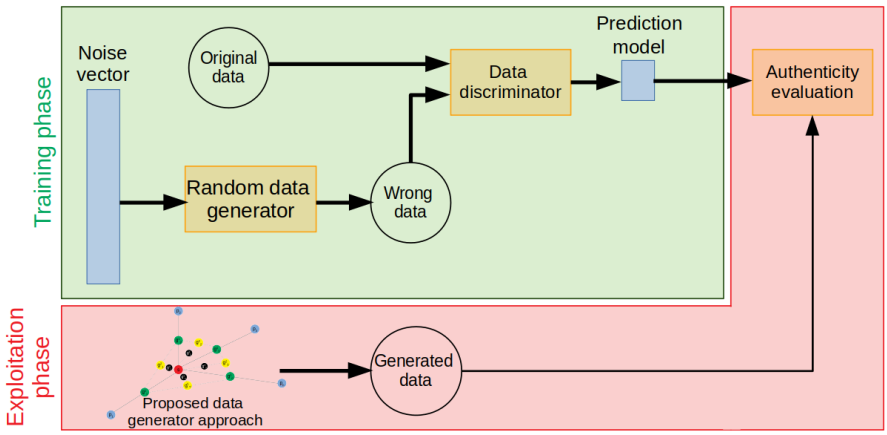


FIGURE 3.5: Illustration of the data discriminator principle.

[Goodfellow et al., 2014]. However, the discriminator in this work is trained to differentiate the original data and simulated data with a high level of performance, unlike that used in the GAN algorithm, which is trained to achieve only 50% of accuracy. To do this, an artificial neural network is trained to classify the data into two classes Original (1) and Simulated (0). This classifier is trained in one real data set and another randomly generated. Although the data's complexity is unknown, we propose to use an evolutionary algorithm [Zemouri et al., 2019] to obtain the best ANN architecture. The Algorithm 2 presents the different stages of the data authenticity assessment process.

Algorithm 2 The process of assessing the generated data authenticity.

Initialization :
$EANN$ = evolutionary Artificial Neural Network;
B = original data set;
F = set of data generated with a random process;
G = set of data generated with the Algorithm 1;
GA = set of generated and authentic data;

Algorithm:
Train $EANN$ to differentiate data from B and F.;
for (x in G) {
if($RNAE(x)$ == authentic) {
$GA = GA + x$; }
}
The data authenticity is evaluated.

Once the data authenticity is evaluated, they can be used in the explanation process. The next paragraph details the proposed semi-global explanation approach.

Step 4: Semi-global explanation of the results

The first step in the explanation process is to use the black box model $b(.)$ to label the newly generated data. This task will identify the behavior of $b(.)$ in the decision surface. In this context, a decision d is a function that serves to differentiate the instances in the decision plane. In 2D, the function used to classify between instances is a row, while the function used to classify instances in 3D is called a plan, just as the function which classifies the point in a higher dimension is called hyperplane (see Figure 3.3 (d)). In general, the equation of the hyperplane in n dimensions can be given by:

$$\alpha^T \times \mathbf{V} + c. \tag{3.4}$$

where c is a constant, $\mathbf{V} = (V_1, \ldots, V_l)$ is the vector of variables and α^T is the leading vector of the decision plane.

The numerical formula for α^T is then determined using the "Soft margin" technique [Shawe-Taylor and Cristianini, 2002]. To do this, we consider the case where there are two classes $C1$ and $C2$ that refer to the healthy and faulty classes which corresponds to a detection problem. A decision d_i for an instance X_i can take two values: $d_1 = -1$ when $\{X_i \in C1\}$ and $d_2 = 1$ if $\{X_i \in C2\}$. Thus, an instance X_i is well explained if the following condition is satisfied:

$$d_i \times (\alpha^T \times X_i + c) \geq 1, \quad d_i \in \{-1, 1\}. \tag{3.5}$$

This condition requires that the decision plan properly explains all decisions. This requirement is, therefore, difficult to meet in reality. For this reason, we suggest allowing some bad explanations in the dataset. To do this, we will grant a constant $\epsilon_i \geq 0$ which for each instance X_i, we have:

$$d_i \times (\alpha^T \times X_i + c) \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0. \tag{3.6}$$

This new constraint makes it possible to accept imperfect classifications. However, the objective is to minimize these imperfect classifications. Thus, the leading vector of the decision plan α^T is determined as follows:

$$\begin{cases} \min \sum_{i=1}^{n} \epsilon_i \\ Constraints: \\ \qquad d_i \times (\alpha^T \times X_i + c) > 1 - \epsilon_i. \\ \qquad \epsilon_i \geq 0. \end{cases} \tag{3.7}$$

The proposed explanation approach allows explaining decisions in a semi-global way for binary classification problems (only two classes). However, multi-class problems can be treated in the same way as binary classification problems. Thus, a multi-class problem can be transformed into a two-class problem by considering only the class of the instance X_i to be explained and the rest of the classes as a single class. In the following section, a validation of the approach is conducted in a real case study.

3.6 SCODER Case study validation (Part 1)

The proposed approach is applied in the SCODER company. This section reports the application steps and the obtained results.

3.6.1 Application and results

The application consists of a stamping line where sheet metal properties are controlled to identify their impact on production performance. Each metal coil is represented by a set of mechanical and chemical proprieties. The class is a binary variable that indicates if the metal coil is suitable for production or not. The proposed framework is used to understand each metal propriety's impact on production performance and identify the optimal sheet-metal characteristics for stable production. In this context, the SCODER's staff have excellent know-how that can be used to guide the knowledge extraction phase. The previously detailed approach is used in order to make the extracted rules transparent. Below, the steps of this application are detailed:

1. Data Preparation The metal coils proprieties are collected and crossed with the machine breakdowns to train a black-box machine learning model.

2. Knowledge integration In this step, and by using the process defined in Section 3.5.1, the operators' know-how is used in different levels:

- Features selection: from the collected data, only five variables are chosen as important. For a confidentiality reason, these variables will be noted $Var_i \ \{i = 1, \ldots, l\}$.
- New variable definition: on the basis of the strength of materials theory, new variables are defined. These variables are: $Var_6 = f(Var_2, Var_3)$, $Var_7 = f(Var_1, Var_2)$ and $Var_8 = f(Var_1, Var_2)$. From the SCODER's staff point of view, these new variables are very important but they create a redundancy in the data. For that, Var_1 and Var_3 are eliminated since they are represented by Var_2, Var_6 and Var_7. Figure 3.6 shows the importance of the features using the Gini criteria [Breiman, 2001] at each step.
- Variables bounds: at this level, human knowledge is used to set the variables' bounds to avoid outliers in the data generation step for the explanation phase.

3. Train $b(.)$ The collected data is used to train a black-box algorithm to solve the problem. In this application, a traditional artificial neural network (ANN) algorithm is used to predict each data instance's class.

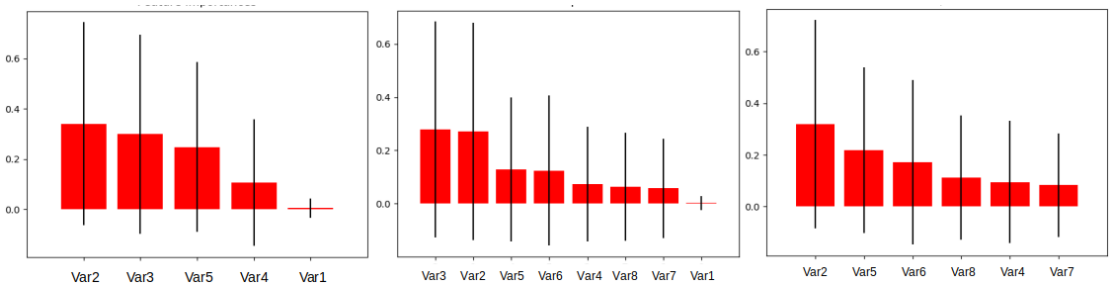


FIGURE 3.6: Features importance at each step of the data generation.

4. Results explanation Using the process described in Section 3.5.2 the learned model by the ANN algorithm is explained for each data instance s. To do this, the **Step 1** of the explanation process is used to identify the decision plan belonging to the instance s. Then, the data generation algorithm described in **Step 2** is applied to generate new data instances in the decision plan. The authenticity of these generated data is assessed using the Algorithm 2 described in **Step 3** of the explanation process. Finally, the explanation process detailed in the **Step 4** is used to provide a semi-global explanation of the decision belonging to an instance s. Figure 3.7 shows an explanation example. This explanation consists of the separation plan which separates the different classes, which is given below:

$$\zeta(b,s) = \begin{cases} C1 \; if \; 0.66 \times Var_2 - 0.26 \times Var_4 + 0.26 \times Var_5 - 0.50 \\ \quad \times Var_6 - 0.19 \times Var_7 - 0.19 \times Var_8 < 233 \\ C2 \; if \; 0.66 \times Var_2 - 0.26 \times Var_4 + 0.26 \times Var_5 - 0.50 \\ \quad \times Var_6 - 0.19 \times Var_7 - 0.19 \times Var_8 \geq 233 \end{cases} \tag{3.8}$$

where $C1$ is the class of good metal coils and $C2$ is the class of bad metal coils.

This explanation is valid to each data point in the neighborhood of the instance s, and it gives greater importance to $Var2$. Thus, the greater this variable, the more suitable the metal of the coil is for production.

3.6.2 Discussion

The proposed framework combines the integration of human knowledge, and the explanation of analysis results in a collaborative approach of PHM. This combination allows to improve the data quality and to enrich the human knowledge base. Also, this framework partly participates in creating a corporate memory that is little developed in the case of SMEs. To improve the proposed approach, we offer a non-exhaustive list of points that should be considered as future work:

- Further formalization of the data quality as a function of the knowledge.

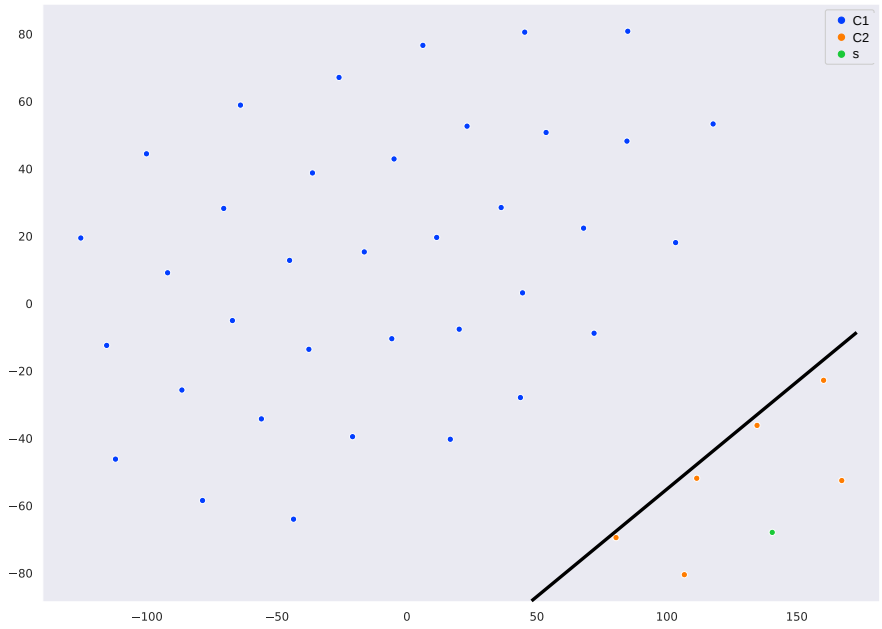


FIGURE 3.7: Explanation results.

- Human knowledge uncertainties quantification and assessment of their impact on data quality improvement results.
- Uncertainties formalization and proposition of correction approaches.

3.7 Conclusion

This chapter focuses on data quality improvement by proposing a new knowledge-based approach for data quality improvement and know-how enrichment. The chapter has been structured around two essential themes: informed and explainable learning.

Firstly, human knowledge types are defined, and their integration modes in the data management process are discussed. This study is valorized in a new knowledge-based data quality improvement framework. Concerning the enrichment of the human knowledge base, the obtained analysis results are explained and used to enrich the human knowledge base. This same knowledge is applied to improve data quality in future analysis tasks. The proposed approach is applied in the SCODER case study in order to validate it and assess its applicability.

The proposed approach allows a better master of the data quality problem in the Industry 4.0 context. Thus, the technological and methodological feasibility of the proposed approach is proved. However, the economic justification is not always apparent,

especially in the case of SMEs with limited financial resources. The next chapter focuses on the data quality economic aspect by optimizing data quality and improving industrial performance.

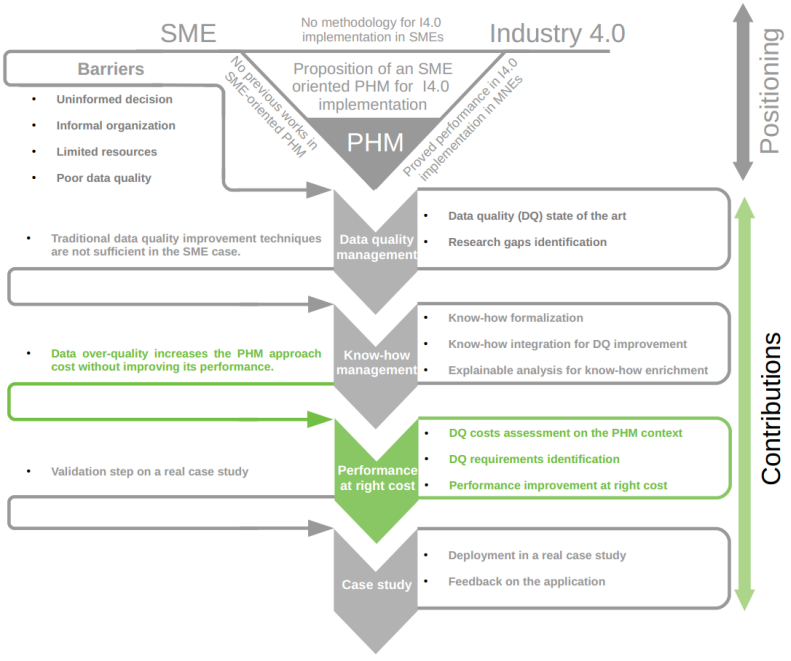
Chapter 4

Data quality optimization for performance improvement

"A theory is a supposition which we hope to be true, a hypothesis is a supposition which we expect to be useful; fictions belong to the realm of art; if made to intrude elsewhere, they become either make-believes or mistakes."

-George Stoney

Graphical abstract.



Contents

Contributions

This chapter deals with the data quality cost optimization. To do, the data quality impact on the fault detection task of the PHM process is formalized. This formalization is used to propose an empiric metric to quantify this impact. Based on this empirical metric, different scenarios for data quality cost optimization are proposed. The proposed approach is applied in the SCODER case study to validate it and assess its applicability.

4.1 Introduction

This chapter addresses the economic aspect of the data quality by proposing to improve them and thus industrial performance at the right cost. The unavailability of financial resources prevents SMEs from installing sophisticated digitization infrastructure, which results in weak data quality. Thus, SMEs fail to use these data to improve their performance. In this chapter, we propose to empirically model the impact of the data quality on industrial performance. These developed models are then used to optimize the data cost while satisfying the desired performance level.

The remainder of this chapter is organized as follows. Section 4.2 discusses the importance of the economic aspect of data quality. In Section 4.3, the relationship between data quality and performance is formalized, and an associated empirical metric is presented. Section 4.4 proposes different approaches to optimize the data quality cost and improve performance at the right cost. These developments are applied in the SCODER case study in Section 4.5. Finally, conclusions are displayed in Sections 4.6.

4.2 Data quality optimization

Data quality, which is the backbone of this thesis work, directly impacts the decisions taken to improve the performance [Omri et al., 2021]. The cleaner the data, the higher the obtained performance. However, the higher the DQ level, the more expensive the acquisition cost [Haug et al., 2011]. This cost is generally expensive for SMEs with limited resources. For that, this cost must be optimized while satisfying an acceptable performance level.

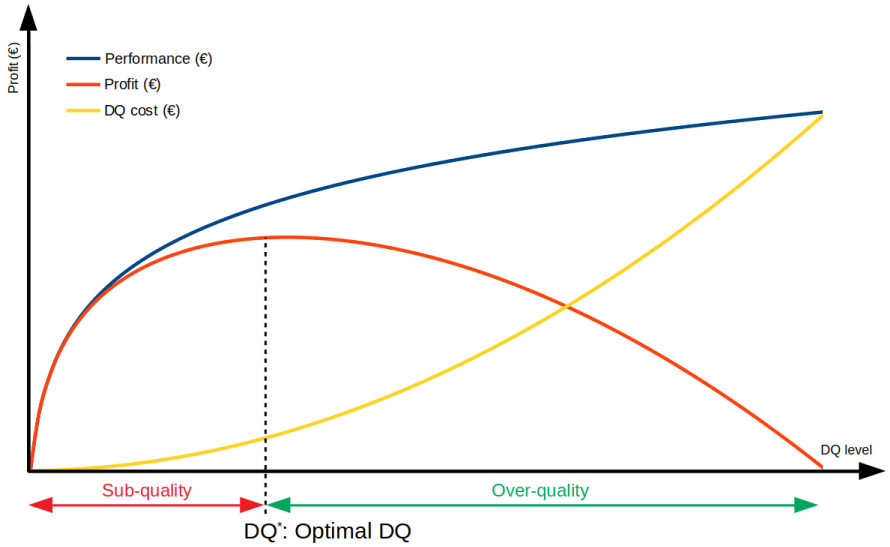


FIGURE 4.1: Profit evolution as a function of data quality. In yellow, the evolution of data quality costs in relation to performed DQ level. The higher the DQ level, the more expensive the acquisition cost. In blue, the performance evolution regarding the used data quality. In red, the profit is calculated as the subtraction of the performance from the DQ cost.

As shown in Figure 4.1, performance income and acquisition costs increase as the level of used data quality increases. Thus, an optimal DQ level (DQ^*) must be identified in order to maximize the performance with a minimum DQ cost [Eppler and Helfert, 2004].

The objective of this work is to improve performance while optimizing costs. Thus, the data acquisition costs must be reduced as much as possible while guaranteeing an acceptable performance level. However, the collected data quality strongly influences the obtained results quality. For this reason, a set of data quality requirements should be defined in order to meet the objectives of a given project. Considering that data quality strongly depends on the used acquisition technologies and the data inventory, it seems interesting to optimize these data qualities to reduce the acquisition costs and thus the

final DQ cost. In the following sections, we formalize the link between data quality and the obtained results while detailing a performance improvement methodology at the right cost.

4.3 Data quality impact on the performance

In this section, we propose defining a set of data quality requirements that should be respected to meet the objectives of a given project. For that, the data quality impact on the PHM results is first formalized then an empirical metric is developed to assess this impact.

4.3.1 Development intuitions and assumptions

As shown in Fig. 4.2, data quality management in the PHM context can be seen from two sides: (1) a straightforward process where data quality is assessed and its suitability for the PHM application is evaluated, and (2) a reverse process where a set of data quality requirements are defined to meet the fixed objectives. In the PHM context, there is little literature that addresses the data quality issue. These works analyze the adequacy of an existing data set to the fixed objectives. This implies that the data acquisition step is carried out in advance. Moreover, these works are based on visualization techniques for data quality assessment without defining a generic metric to quantify data quality and its impact on PHM results. We are here interested in defining a generic metric that allows the understanding and quantification of the data quality impact on PHM tasks concerning each task's expected performance before installing the data acquisition system. Recall that the main PHM tasks are fault detection, diagnosis, and degradation prediction. The fault detection task is the first one on the PHM process [Jia et al., 2017] and is considered in the rest of this study.

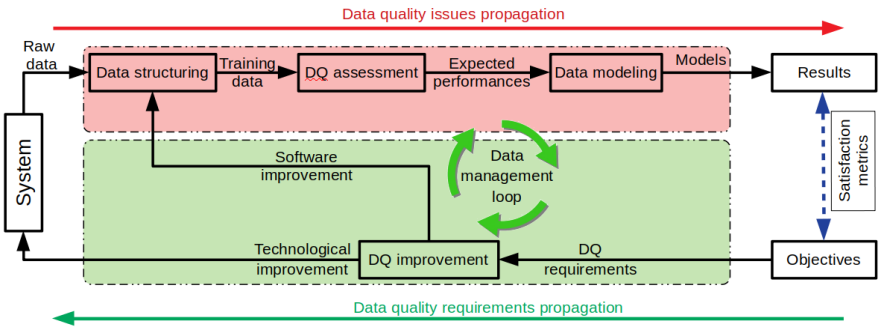


FIGURE 4.2: Data management process [Omri et al., 2021]. In red, the straightforward process consists of evaluating the suitability of the used data to the fixed objectives. In contrast, the inverse process (in green) aims to set a data quality requirement that should be respected to satisfy the objectives. For that, data quality improvement actions are proposed at the system level and the data level.

In [Roy and Dey, 2018], the authors specify that fault detectability can be divided into two notions: *(i) Intrinsic detectability* and *(ii) Performance-based detectability*. The intrinsic notion refers to the system's anomalies signature without dependence on the used fault detection technique. This fits with the system's intrinsic propriety such as controllability and observability [Ding, 2008]. On the other side, performance-based fault detectability is defined according to the used fault detection algorithm, and it refers to the ability of this algorithm to detect anomalies [Roy and Dey, 2018]. As shown in Fig. 4.3, many factors can affect the fault detection task. These factors can be related to the used detection algorithm's performance, the data quality issues, or the system observability. For the first possibility, many sophisticated algorithms have been proposed to deal with fault detection with impressive performances. However, suppose the used data do not describe the studied system. In that case, it is not necessary to develop a sophisticated algorithm to solve the problem because it is impossible to meet the objectives due to the data's inadequacy [Omri et al., 2021].

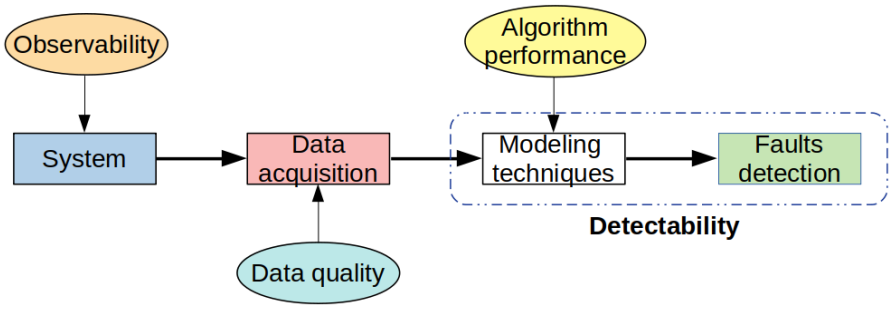


FIGURE 4.3: Factors that impact the detectability accuracy [Omri et al., 2021].

In this section, we propose to formalize the data quality impact on the fault detection task. Some assumptions are made for our study:

(A1) The identified variables in the data inventory step and the operators know-how provide a complete description of the system Σ.

(A2) $\sum$ is observed during a sufficient horizon of time to collect the needed data.

(A3) The used detection algorithms are all able to perform equal results.

(A4) The detectability task is done in a supervised mode.

To sum up, this section aims to quantify detectability for fully observable systems and define data quality requirements concerning the expected detection results.

4.3.2 Data quality problem formulation

Intrinsic detectability refers to the system's anomalies signature without any dependence on the used fault detection technique. This fits with the observability O as a system's intrinsic propriety. On the other side, the performance-based fault detectability is defined according to the used fault detection algorithm, and it refers to the ability of this algorithm to detect anomalies. However, an algorithm's ability to detect anomalies can be a result of its intrinsic performance P and the quality of the used dataset Q. Thus, the detectability of a system $\sum$ can be expressed as a function of the observability, the data quality, and the performance of the used detection algorithm:

$$Det = f(O, Q, P). \tag{4.1}$$

where $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ describes the link that exists between O, P, Q and the detectability.

As detailed above, we are only interested in studying the data quality impact by considering (A1-A4). Thus, the detectability can be expressed as a function of the data quality:

$$Det = f(Q). \tag{4.2}$$

Data quality stands out as one of the essential criteria since it impacts the used detectability algorithm's performance. We have to point out that global data quality issues belong to the dataset (i.e., imbalanced data) and other local issues pertaining to variables (i.e., missing data or noisy data). Thus, each type of data quality acts differently on the detection task. The global quality issues (GQ) have an iso-impact on each feature X_i regardless of its local quality problems (LQ_i) as shown below:

$$Q_i = GQ \times LQ_i \ , \forall \ i \in \mathbb{N}. \tag{4.3}$$

The GQ is the quality issues that concern the whole dataset, which is described by:

$$GQ = \prod_{j=1}^{m} GQ_j. \tag{4.4}$$

where m is the number of the considered global data quality problems GQ_i.

As for the local quality issues, their impacts differ from a variable X_i to another. The link between the local quality of a feature X_i with the different l quality problem that concern this feature is described by:

$$LQ_i = \sum_{k=1}^{l} LQ_{ik}, \forall i \in \mathbb{N}. \tag{4.5}$$

where LQ_{ik} is a local quality that depends on the quality characteristic q_{ik} and the feature importance weight w_i. Thus, LQ_{ik} is a complex function that connects these variables given by:

$$LQ_{ik} = g(w_i, q_{ik}). \tag{4.6}$$

where $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ that describes the local quality of a feature X_i regarding a data problem k.

According to (4.3), (4.4) and (4.5), the quality of a feature X_i can be written by:

$$Q_i = \prod_{j=1}^{m} GQ_j \times \sum_{k=1}^{l} LQ_{ik}. \tag{4.7}$$

Using (4.7) and (4.6), the quality of the recorded dataset can be explained concerning the feature importance weights by:

$$Q = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} \Big[\prod_{j=1}^{m} GQ_j \times \sum_{k=1}^{l} g(w_i, q_{ik}) \Big]. \tag{4.8}$$

Hence, referring to the development detailed in (4.7) and (4.8), the detectability metric can be written as follows:

$$Det = \prod_{j=1}^{m} GQ_j \times \sum_{i=1}^{n} \sum_{k=1}^{l} g(w_i, q_{ik}). \tag{4.9}$$

4.3.3 The empirical data quality model

This section presents an empirical development of the detectability metric using the previously detailed problem formulation in Section 4.3. Thus, we propose to estimate the global and local data quality as proposed in (4.9). To do this, it seems logical to estimate the parameters w_i from the ability of features to detect the system's abnormal mode. However, it may be more challenging to estimate the global and local quality functions. Thus, we propose to estimate these elements empirically.

The features importance w_i are essential parameters for any data analysis task that cannot be overlooked or marginalized. In this context, two solutions arise to define feature importance: (i) based on human expertise or (ii) based on manually collected data. The first solution seems to be easier, faster, less expensive, but imprecise. Human expertise is limited in the case of a complex problem. Since this work results from a practical approach, the second solution is adopted due to its precision. Data samples are collected carefully and manually, and they are used to preliminary analyze the data and quantify the importance of each feature. We refer here to the feature importance conducted implicitly by the *Random Forest* classifier based on the "Gini importance" [Breiman, 2001]. According to this method, the importance of a feature X_i is computed by the sum of all impurity decrease measures of all nodes in the forest at which a split on X_i has been conducted and normalized by the number of trees [Nembrini et al., 2018]. The impurity for a tree t is usually computed by the Gini impurity given below:

$$G^t(X_i) = \sum_{K=1}^{Category(X_i)} p_a(K) \times G(K). \tag{4.10}$$

where X_i is the feature, $p_a(K)$ is the fraction of category K in a feature X_i and $G(K) = \sum_{a=1}^{C} p_a(K) \times (1 - p_a(K))$ is the gini index of a category K.

Then, the feature importance is obtained as follow.

$$w_i = \frac{1}{n_{tree}}[1 - \sum_{t=1}^{n_{tree}} G^t(X_i)] \tag{4.11}$$

where n_{tree} is the number of trees.

We then turn to estimate the local and global quality functions. For that purpose, we considered ten datasets (real and simulated datasets), and we tested the most used fault detection techniques to study their behavior regarding the data problems. Table 4.1 presents the training datasets used to study the behavior of the most used fault detection algorithms regarding data quality problems. In this study, the used algorithms include:

TABLE 4.1: Details of the training datasets.

Dataset	Number of features	Number of instances	Application domain	Reference
Credit card	24	30000	Credit card default	[Yeh and Lien, 2009]
DBWorld e-mails	4702	64	Announces detection	[Filannino, 2011]
BCWD	10	699	Breast cancer detection	[Wolberg and Mangasarian, 1990]
Car Evaluation	6	1728	Car safety detection	[Bohanec and Rajkovic, 1988]
Balloons	4	16	Cognitive psychology	[Ross et al., 1990]
Audit	18	777	Fraudulent firm detection	[Hooda et al., 2018]
Dataset 1	5	10000	Artificial data	-
Dataset 2	10	10000	Artificial data	-
Dataset 3	15	10000	Artificial data	-
Dataset 4	20	10000	Artificial data	-

- Artificial neural network (ANN): Given a set of features and a target, an ANN can learn a non-linear function that can be used for classification or regression. ANN is different from logistic regression because, between the input layer and the output layer, it can be one or more non-linear layers, called hidden layers [Zemouri et al., 2019].
- Decision tree (DT): The main idea of the DT algorithm is to learn from the data to create simple inferred rules that will be used to segment the data and make predictions [Tso and Yau, 2007].
- Support vector machine (SVM): The SVM aims to find a separating hyperplane that separates the different classes. The hyperplane that reduces the number of wrongly classified samples in the training phase is called Optimal Separating Hyperplane (OSH).

- K-nearest neighbors (KNN): The KNN classifier consists of predicting the class of a new point based on the classes of the k closest instances to this later [Khanzadeh et al., 2018].
- Naive Bayes (NB): The NB algorithm is based on coupling the *Bayes theorem* with the *Naive* hypothesis of conditional independence between every pair of features given the value of the class variable. More details about this technique are presented in this work [Rish et al., 2001].

Recall that this work aims to quantify DQ impact on the most commonly used fault detection algorithms. To do so, we briefly present these algorithms, and we encourage readers to consult the mentioned references for more details on these techniques.

Before detailing the obtained data quality models, this paragraph describes the injection of data quality problems in the training datasets. For the missing data problem, original values are replaced by the value 0. As mentioned above, a value is considered noisy only if it impacts the detection result. However, variables are dependent, which means that a variable X_i can be considered noisy or not regarding other features' accuracy. For this, we randomly add noises ϵ_i to each feature X_i (such as $-mean(X_i) \leq \epsilon_i \leq mean(X_i)$) and we evaluate if these noises affect the detection result. Then we define the noise threshold for each feature X_i as $mean(\epsilon_i)$. Thus, added noises are superior to these thresholds. For the imbalanced data, the instance number of the faulty class is modified to create a between-class imbalance. More than 10^5 simulations have been carried out with different quality configurations. For each configuration, the data detectability is assessed. The overall mean of these simulation results is then used to develop a global detectability model considering each data quality issue (i.e., Imbalanced, missing, and noisy issues). The obtained models are detailed below.

- **Imbalanced data model:** Numerical simulations performed on the different datasets have shown that detectability increases exponentially in function of the imbalanced data ratio. This evolution is illustrated in Fig. 4.4 and shows that the imbalanced data quality issue has no impact on the detectability result if its ratio is greater than 50%. The global quality, defined in (4.4) for $m = 1$ (since we only consider the imbalanced data as a global quality issue), is then given by:

$$GQ(q_{Im}) = 1 - 0.52 \times e^{-0.07 \times q_{Im}} \tag{4.12}$$

  where q_{Im} is the imbalanced data ratio defined in Equation (2.2) in Chapter 2.

- **Missing data model:** Fig. 4.5 displays the detectability evolution as a function of the missing data ratio per feature. These results show that the detectability decreases as the missing data ratio increases. The impact of this problem is more evident when the missing ratio exceeds 40%.

  The local quality, related to the missing data ratio q_{i1} of a feature X_i, depends on the evolution function of Fig. 4.5 and multiplied by the term $\frac{w_i}{w_{i_{min}}}$ to describe the detectability evolution function of feature importance.

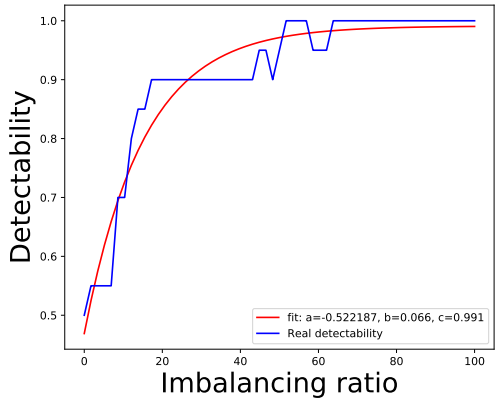


FIGURE 4.4: Detectability evolution as a function of the imbalanced data ratio.

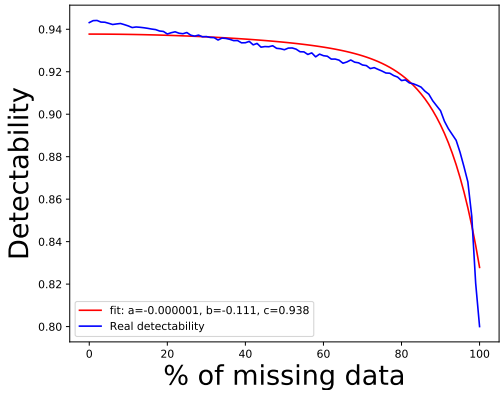


FIGURE 4.5: Detectability evolution function of the missing data ratio per feature.

$$LQ_{i1} = g(w_i, q_{i1}) = \frac{w_i}{w_{i_{min}}}[1 - 2.10^{-6} \times (q_{i1}^2 + e^{0.11 \times q_{i1}})] \tag{4.13}$$

where q_{i1} is the missing data ratio defined in Equation (2.5) in Chapter 2.

- **Noisy data model:** Fig. 4.6 displays the detectability evolution function of the noisy data ratio per feature. These results show that detectability decreases when the noisy data ratio increases. Like the missing data issue, this problem's impact is more evident when the noisy ratio exceeds 20%.

  Therefore, the local quality related to the noisy data is given by

$$LQ_{i2} = g(w_i, q_{i2}) = \frac{w_i}{w_{i_{min}}}[1 - 10^{-6} \times (q_{i2}^2 + e^{0.07 \times q_{i2}})]. \tag{4.14}$$

  where q_{i2} is the noisy data ratio defined in Equation (2.4) in Chapter 2.

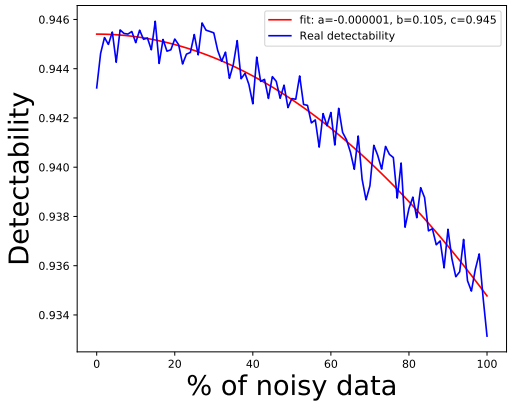


FIGURE 4.6: Detectability evolution function of the noisy data ratio per feature.

- **Final detectability model:** The previously detailed models details one global data quality issue (imbalanced data) and two local data quality problem (noisy and missing data). By substituting them into the detectability model proposed in (4.9), the final detectability model can be defined as follows:

$$\begin{aligned} Det &= GQ(q_{Im}) \times \sum_{i=1}^{n} [g(wi,q_{i1}) + g(wi,q_{i2})] \\ &= (1 - 0.52 \times e^{-0.07 \times q_{Im}}) \times \sum_{i=1}^{n} \frac{w_i}{w_{i_{min}}} [2 - 2.10^{-6} \times (q_{i1}^2 + e^{0.11 \times q_{i1}}) \\ &- 10^{-6} \times (q_{i2}^2 + e^{0.07 \times q_{i2}})]. \end{aligned} \tag{4.15}$$

To better understand the impact of the studied data quality issues on the fault detection task, Fig. 4.7 displays the detectability map in function of the fundamental data quality issues. It is shown that the imbalanced data ratio has a fatal impact on detectability when it is less than 20%. Moreover, the missing data ratio has a significant impact when it is greater than 80%. One should note that the proposed detectability metric is derived from an accurate understanding of the data's behavior and the related data quality problems.

Finally, we come to assess the accuracy of the developed model. A set of numerical simulations is used to validate the detectability model. Thus, the previously used fault detection algorithms are tested to define their behavior regarding the data problems. Table 4.2 shows the results of the validation steps. For each dataset, 500 data quality configurations are tested. Results show that the developed model can predict the general evolution of detectability as a function of the used data quality.

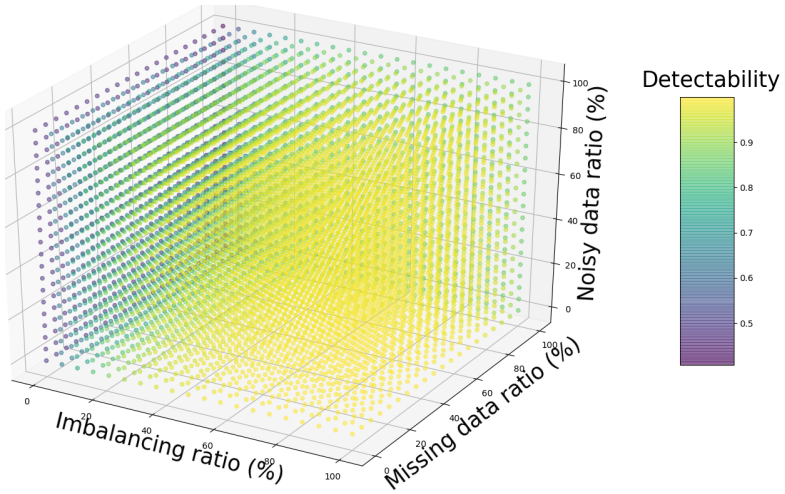


FIGURE 4.7: Detectability map in function of the basic data quality issues.

TABLE 4.2: Results of the model validation step.

Dataset	Reference	# of features	# of instances	Application domain	RMSE
Diagnosis	[Czerniak and Zarzycki, 2003]	6	120	Inflammation detection	0.05
Spam	[Wang and Witten, 2002]	57	4601	Spam detection	0.09
Blood Transfusion	[Yeh et al., 2009]	4	748	Blood Transfusion	0.08
Caesarian	[Amin and Ali, 2018]	5	80	Caesarian Section detection	0.04
Cryotherapy	[Khozeimeh et al., 2017]	6	90	Wart treatment	0.09

Detectability is predicted with a root-mean-square error (RMSE) less than 0.1. We can affirm that the developed model can quantify the data quality impact on the PHM results with an acceptable performance level. Recall that the data quality depends on the used acquisition technologies. The following section proposes to use the developed data quality model to define the required data quality for a fixed objective and thus, optimize the data quality cost.

4.4 Performance improvement at right cost

In this section, we propose to study the different cost elements of DQ in detail to optimize the acquisition cost. As shown in Figure 4.8, optimization can take several forms, such as (i) Minimizing the DQ cost and (ii) Maximizing performance.

The main objective of this part is to improve industrial performance at the right cost. However, this objective requires the existence of specific technologies and the resulting cost. On the other hand, this cost may be inappropriate for SMEs with limited financial resources. Thus, we propose to allow SMEs to access revealing technologies of Industry 4.0 with limited budgets. Thus, we formalize these objectives in two scenarios.

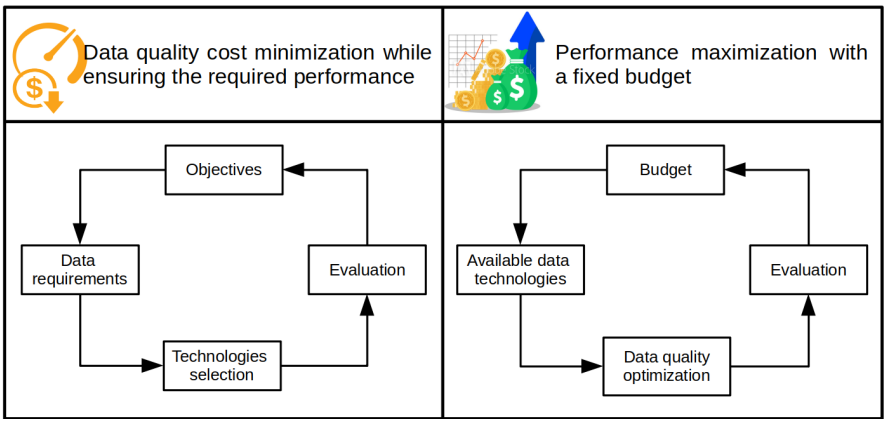


FIGURE 4.8: Strategies for optimizing the DQ cost.

- The first scenario requires reaching a defined performance level while minimizing the data quality cost. To do, the developed data quality models are used to define the data quality requirements to meet the fixed objectives. Based on these requirements, the necessary acquisition technologies are chosen according to their cost.
- The second scenario assumes that the available budget to install the Industry 4.0 concept is fixed. Thus, the objective, in this case, is to maximize performance while respecting the available budget. To do, an inventory of existing data technologies is carried out to optimize the variables to be collected and their level of quality. This optimization is done in order to maximize the resulting performance.

In the following paragraphs, we detail these objectives by formalizing each scenario.

4.4.1 Data quality cost minimization

In this paragraph, we propose to improve industrial performance while optimizing the DQ cost. Based on the developed data quality models, we propose here to optimize data acquisition costs while guaranteeing the satisfaction of the fixed objectives.

Variables Let q_{Im} be the imbalanced data ratio, q_{i1} the missing data ratio and q_{i2} the noisy data ratio. As proved in the previous section, these data qualities strongly impact the obtained performance. Different costs are allocated to guarantee each quality level such as CD is the cost of negative detection (false alarm), CM_j and CN_j are respectively the costs to guarantee a missing data ratio q_{i1} and a noisy data ratio q_{i2} and n is the variable number.

Constraints The constraint, detailed in the equation (4.16), ensures the satisfaction of a required level of performance $Perf$:

$$Perf \leq GQ(q_{Im}) \times \sum_{i=1}^{n}[g(wi,q_{i1}) + g(wi,q_{i2})] \leq 100. \tag{4.16}$$

Objective function The objective of this model is to minimize the overall data quality cost:

$$\begin{cases} \min \ \big[CI \times q_{Im} + CD \times (100 - GQ(q_{Im}) \times \sum_{i=1}^{n}[g(wi,q_{i1}) + g(wi,q_{i2})]) \\ \quad + \sum_{i=1}^{n} CM_j \times (100 - q_{i1}) + \sum_{i=1}^{n} CN_i \times (100 - q_{i2})\big]. \end{cases} \tag{4.17}$$

Scenario formalization The following mathematical problem can be used to optimize the objective defined by this scenario, which is the minimization of data quality cost while ensuring a defined performance level:

$$\begin{cases} \min \ \big[CI \times q_{Im} + CD \times (100 - GQ(q_{Im}) \times \sum_{i=1}^{n}[g(wi,q_{i1}) + g(wi,q_{i2})]) \\ \quad + \sum_{i=1}^{n} CM_i \times (100 - q_{i1}) + \sum_{i=1}^{n} CN_i \times (100 - q_{i2})\big] \\ subject\ to: \\ \qquad Perf \leq GQ(q_{Im}) \times \sum_{i=1}^{n}[g(wi,q_{i1}) + g(wi,q_{i2})] \leq 100 \\ \qquad 0 \leq q_{i1} \leq 100, \ for\ i = 1,\ldots,n \\ \qquad 0 \leq q_{i2} \leq 100, \ for\ i = 1,2,...,n \\ \qquad 0 \leq q_{Im} \leq 100 \end{cases} \tag{4.18}$$

where CD is the cost of negative detection (false alarm), CM_i and CN_i are respectively the costs to ensure a missing data ratio q_{i1} and a noisy data ratio q_{i2} and n is the variables number.

4.4.2 Maximize industrial performance

We propose in this paragraph to optimize the industrial performance while respecting an available budget. Based on the developed data quality models, we propose improving the data analysis results with a fixed budget.

Variables Let q_{Im} be the imbalanced data ratio, q_{i1} the missing data ratio and q_{i2} the noisy data ratio. As proved in the previous section, these data qualities strongly impact the obtained performance. Different costs are allocated to guarantee each quality level such as CD is the cost of negative detection (false alarm), CM_j and CN_j are respectively the costs to guarantee a missing data ratio q_{i1} and a noisy data ratio q_{i2} and n is the variable number.

Constraints The constraint, detailed in the equation (4.19), ensures that the available budget is respected:

$$CI \times q_{Im} + \sum_{i=1}^{n} CM_j \times (100 - q_{i1}) + \sum_{i=1}^{n} CN_i \times (100 - q_{i2}) \leq Budget. \tag{4.19}$$

Objective function The objective is to maximize the performance of the prediction model:

$$\max\ GQ(q_{Im}) \times \sum_{i=1}^{n} [g(wi,q_{i1}) + g(wi,q_{i2})]. \tag{4.20}$$

Scenario formalization The following mathematical problem can be used to optimize the objective defined by this scenario, which concerns the performance maximization while respecting a fixed budget:

$$\begin{cases} \max\ GQ(q_{Im}) \times \sum_{i=1}^{n} [g(wi,q_{i1}) + g(wi,q_{i2})] \\ subject\ to: \\ \qquad CI \times q_{Im} + \sum_{i=1}^{n} CM_j \times (100 - q_{i1}) + \sum_{i=1}^{n} CN_i \times (100 - q_{i2}) \leq Budget \\ \qquad 0 \leq q_{i1} \leq 100,\ for\ i = 1, \ldots, n \\ \qquad 0 \leq q_{i2} \leq 100,\ for\ i = 1, \ldots, n \\ \qquad 0 \leq q_{Im} \leq 100 \end{cases} \tag{4.21}$$

where CM_i and CN_i are respectively the costs to ensure a missing data ratio q_{i1} and a noisy data ratio q_{i2} and n is the variables number.

The following section presents a practical application of this work in the SCODER case study.

4.5 SCODER case study validation (Part 2)

We here consider the SCODER case study as a real application of the proposed approach. This section reports the application steps and the obtained results.

4.5.1 Application and results

The objective is to ensure stable production by reducing machine failures and improving productivity and quality. The production performance is affected by the used metal coil characteristics. For that purpose, a PHM study is conducted to determine if sheet metal is suitable for production or not according to the quantity of non-conform parts produced. This study is based on the coil's characteristics, the caused press breakdowns, and the quality rate of the products fabricated from the sheet metal coil. The aim here

is to identify each metal coil's suitability with 80% as the minimum rate of performance. Algorithm 3 presents the steps to be followed.

Algorithm 3 Data quality management algorithm

Step 1: Identify the problem and understand it.
Step 2: Collect some samples that can describe the problem.
Step 3: Compute the features importance and identify the most important ones.
Step 4: Apply the data quality models and identify the requirements according to objective.
Step 5: Install the data acquisition system.
Step 6: Control and improve the results.

A data inventory is first conducted to collect all the data that can be useful for the project. Then, 24 variables are identified, which consist of 12 metal proprieties, six types of machine breakdown, and six kinds of product's non-conformity. Samples of these data are collected carefully and manually, and they are used to preliminary analyze the data and quantify the importance of each feature. Only five features are identified as pertinent for the study. Thus, the rest of the features are eliminated, and the rest of the study is based on these five variables. Table 4.3 shows the importance of the features from the used data subset. It is proven that the 5^{th} variable is the most important one to identify the capacity of the used coil to produce good quality parts.

TABLE 4.3: Features importance in the SCODER case study.

Variable	Var_1	Var_2	Var_3	Var_4	Var_5
w_i	0.11	0.09	0.14	0.09	0.57

Once the features are identified, the data quality issues for the SCODER dataset are analyzed (see Fig. 4.9). Results show that it is authorized to have an imbalanced data ratio greater than 50%. Moreover, a percentage less than 20% and 30% of noisy data and missing data, respectively, has no impact on the detectability results.

A simple technique to set data quality requirements to satisfy the fixed objectives is to use these thresholds ($q_{Im} \geq 50\%$, $q_{i1} \leq 30\%$ and $q_{i2} \leq 20\%$) to guide the PHM implementation for the SCODER case study. However, this solution does not consider the cost and the time to guarantee these data quality levels.

For the SCODER case study, advanced sensing technologies are required to ensure a high data quality level for the fifth variable. As for the other variables, it can be done quickly. Besides, it takes much time to have a balanced dataset. For that reason, a high cost is allocated to the imbalanced data ratio without forgetting the high cost of a negative detection. Table 4.4 shows the magnitude of these costs.

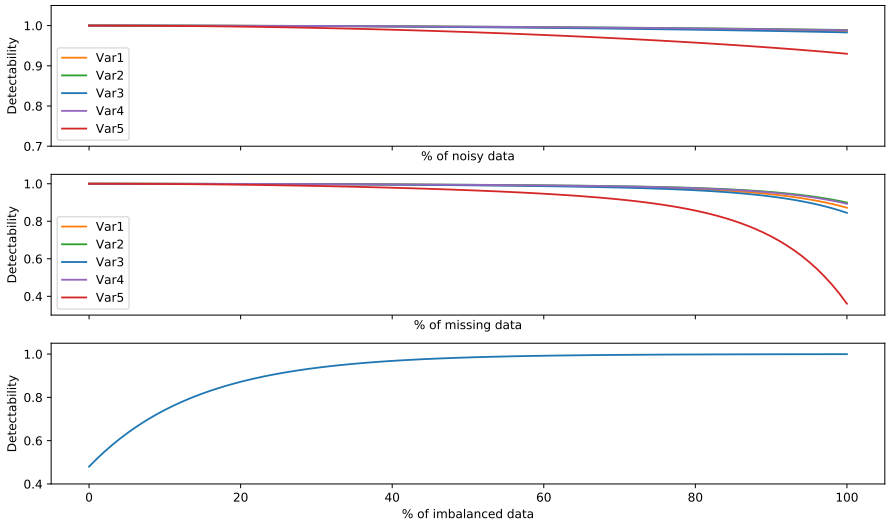


FIGURE 4.9: Synthesis in the SCODER case study.

TABLE 4.4: Magnitude of different costs for the SCODER application.

CD	CI	CM_1	CM_2	CM_3	CM_4	CM_5	CN_1	CN_2	CN_3	CN_4	CN_5
10	8	1	1	1	1	3	1	1	1	1	3

We propose applying the first optimization scenario defined in Section 4.4.1 following the SCODER proposition. This scenario aims to optimize data acquisition costs while guaranteeing the fixed objectives' satisfaction (80% of detectability). Newton's optimization technique [Fischer, 1992] is used to minimize the cost function given in (4.18) and identify the requirements according to SCODER objective. The results of this application are given in Table 4.5.

TABLE 4.5: Data quality requirements for the SCODER application.

Det	q_{Im}	q_{11}	q_{21}	q_{31}	q_{41}	q_{51}	q_{12}	q_{22}	q_{32}	q_{42}	q_{52}
90%	30%	25%	60%	6%	28%	29%	53%	36%	77%	71%	67%

As a matter of fact, it is allowed to have 29% of missing data for var_5 and up to 60% for some other variables. For the noisy data, the percentages are between 36% and 77%. According to the developed data quality model, this configuration results in a detectability of 90%, which satisfies the fixed objective. These requirements are respected during the installation of the data acquisition system. The expected installation cost is 11.05 MU which is optimized for SMEs with limited resources. Figure 4.10 shows the SCODER data acquisition system, which is based on a set of tablets to collect data throughout the production chain.

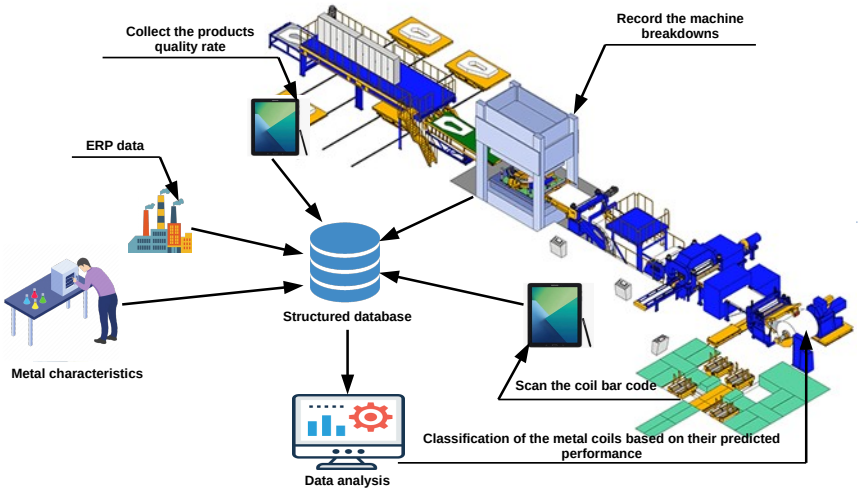


FIGURE 4.10: Details of the SCODER case study.

The metal coils' properties are tested in a specialized test station where the previously defined requirements are met. The dataset was collected in more than six months, during which the data quality has evolved to meet the defined requirements. Table 4.6 displays the evolution of the expected detectability versus the real one using the previously detailed detection algorithms. The results show that the developed model can predict the general evolution of the detectability as a function of the used data quality.

TABLE 4.6: Evolution of the expected detectability versus the real one.

Month	Requirements							Real detectability (%)					Expected detectability (%)
	Imbalanced (%)		Var1	Var2	Var3	Var4	Var5	DT	SVM	ANN	KNN	NB	
M1	5	Missing (%)	94	79	48	83	66	45	**50**	**50**	**50**	**50**	50
		Noisy (%)	0	3	9	0	5						
M2	11	Missing (%)	55	84	65	45	61	80	50	50	**65**	80	67
		Noisy (%)	5	1	4	10	8						
M3	18	Missing (%)	29	53	49	5	55	95	55	**80**	65	50	79
		Noisy (%)	18	18	8	27	21						
M4	25	Missing (%)	31	48	15	47	4	**90**	55	75	65	60	87
		Noisy (%)	13	1	10	9	23						
M5	33	Missing (%)	35	45	17	54	32	95	65	**90**	75	50	92
		Noisy (%)	21	23	24	19	18						
M6	40	Missing (%)	32	41	2	48	28	**95**	70	90	85	90	94
		Noisy (%)	20	10	51	14	19						

4.5.2 Discussion

This work proposes a new model to assess data quality and quantify their impact on the PHM process's fault detection task. This model allows the definition of a set of data quality requirements to satisfy a fixed objective regarding the fault detection task. The algorithm 3 details the various steps to assess data's suitability to the detection task. It should be noted that estimating the importance of features is a difficult task that significantly impacts the developed data quality model's accuracy. In this work, we adopted a solution based on data samples collected manually and carefully. However,

human expertise can be used to accomplish this task. In both cases, the task remains challenging, but from the authors' point of view, the data quality cannot be represented independently of these parameters. Thus, the limitations resulting from the estimation of features' importance should be considered further in future works.

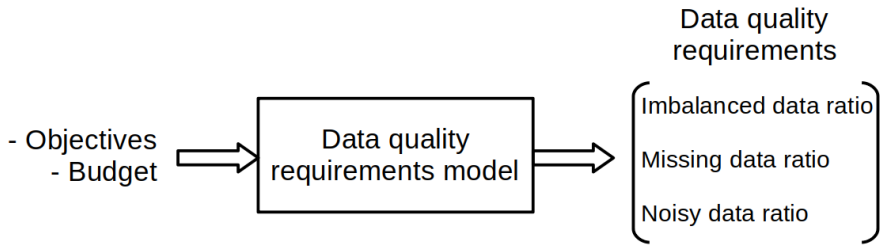


FIGURE 4.11: Identification of data quality requirements based on set objectives and available budget.

As shown in Fig. 4.11, it is possible to formulate a clear idea of the data requirements to be satisfied from the set objectives and the available project budget. These requirements represent the needed data quality ratios, and they can be extended to cover the data storage hubs and data analysis tools. Although the developed models concern only the fault detection task, the same methodology can be used to develop other models for data diagnosability and trendability, which allows covering all the PHM process. Thus, a temporal and technological boundary can be affected to each PHM project. In this way, the PHM strategy cost can be estimated and optimized, which is an understudied topic [Omri et al., 2020].

4.6 Conclusion

This chapter focuses on the economic aspect of data quality by optimizing them for industrial performance improvement at the right cost.

We started this chapter by discussing the data quality economic aspect and its impact on the generated benefits. It is essential to have high data quality. However, data over-quality results in additional costs that can reduce the overall benefits.

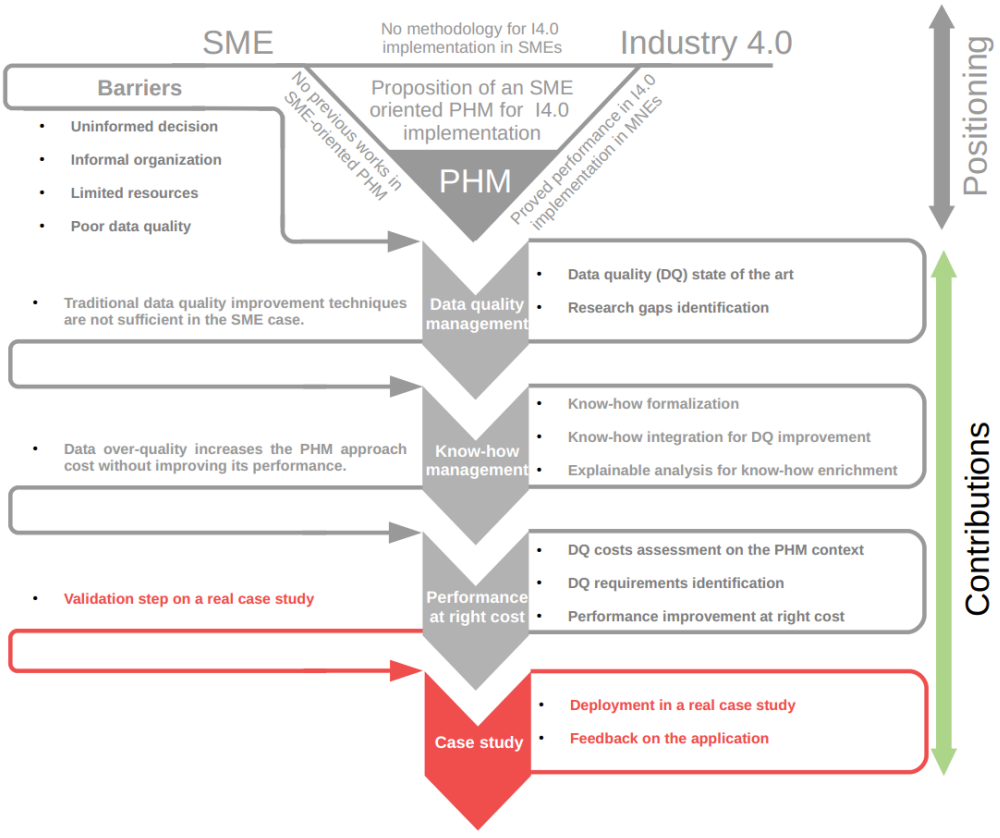
Concerning the optimization of the data quality cost, this task has been done in two steps: assessing the data quality impact on the obtained performance and optimizing the data acquisition cost by identifying data quality requirements. Indeed, data quality depends on used data acquisition technologies. Therefore, the optimization of data quality will allow the appropriate acquisition technologies to optimize the PHM cost. This study provided a first empirical model for data quality requirements identification for PHM applications.

This work was applied in the SCODER company to optimize the data acquisition infrastructure. The obtained results are encouraging and prove the applicability of this work in the industrial domain. Besides, this work can be considered a first step to evaluating the data suitability for a defined PHM project. Further work should be developed to define a technical protocol for data quality evaluation and improvements in a PHM context. This may allow to reduce the time and the cost of data processing and improve decision accuracy.

Chapter 5

SCODER case study validation based on an implemented software

Graphical abstract.



Contents

5.1 Introduction

The developed approaches in this thesis are gradually validated at the end of each chapter. Moreover, a global validation of these works is proposed in this chapter. Indeed, the developments carried out in the previous chapters are formalized in a general methodology for the implementation of Industry 4.0 technologies within SMEs using an adapted PHM approach. This methodology is based on three main pillars: data management, knowledge management, and performance improvement. The conducted application shows good potential and meets the initial objectives of the SCODER company. To facilitate the implementation of this methodology within the SCODER company, all the steps of this methodology have been encapsulated in a SCODER Data System software (DS2). Thus, the various functionalities of the DS2 software are presented. Indeed, its main interfaces are detailed through a real case study of metal coils assignment optimization and following the proposed methodology's three pillars.

The remainder of this chapter is organized as follows. Section 5.2 presents the addressed problem to validate the proposed approach. This approach is encapsulated in the DS2 software, which is presented in Section 5.3. Section 5.4 details the results of this application. Finally, conclusions are displayed in Section 5.5.

5.2 General presentation of the SCODER case study

As mentioned in the chapter 1, this thesis work is applied in the SCODER company to optimize the metal coils assignment. This section proposes to detail the studied problem and the proposed solution.

5.2.1 Metal coils assignment optimization: problem description

The addressed problem concerns developing a framework for optimizing stamping activities in the SCODER company (see Figure 5.1). Indeed, operators notice that production performance depends on the used sheet metal coils. Also, they must find the right die setting for each type of coils. The used strategies to assign coils to the presses or set up the dies are limited in the workers' know-how without any documentation to generalize these skills. To this end, the approach developed in this thesis is applied to understand the problem better and standardize the required actions to maintain a high production performance.

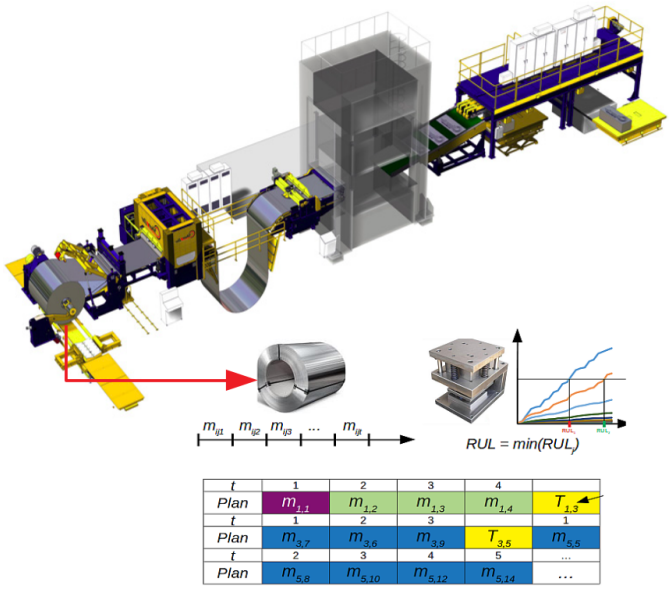


FIGURE 5.1: Coils assignment problem description Where i is the die number, j is the coil reference, t is the period time and m_{ijt} refers to the coil j is affected to the die i in the period t.

In reality, the problem is more complex. SCODER workers are sometimes forced to assign a coil while knowing that this coil will pose many problems during the production process. For this, they try to allocate only small amounts of "bad" coils among many "good" ones to use the wrong stock and reduce problems at the same time. But this is not always possible because there is a significant change in the supply of raw material. Besides, the number of maintenance technicians is higher during the day than at night. Thus, it will be more interesting to allocate the "good" materials only during the night to minimize the problems and ensure the production's continuity. But at the same time, the issues of the day have to be distributed among different machines and at different time intervals to avoid production cuts. In summary, the objectives of this study are to:

- Predict production performance from the characteristics of the used die and metal coil.
- Identify the correct configuration for the die.
- Find the best metal coils sequences.
- Optimize the assignment of these sequences while respecting the constraints mentioned above.

5.2.2 Metal coils assignment optimization: proposed solution

This application aims to develop a framework for optimizing stamping activities in the SCODER company. However, this problem is difficult to solve with exact techniques (i.e., linear programming) given the uncertainty it presents. Indeed, the impact of a metal coil, the die, and the machine's health states are uncertain. Thus, two possibilities of resolution arise: (i) simplify the problem via assumptions and solve it or (ii) solve the problem via case-based reasoning (CBR) approach [Leake, 1996].

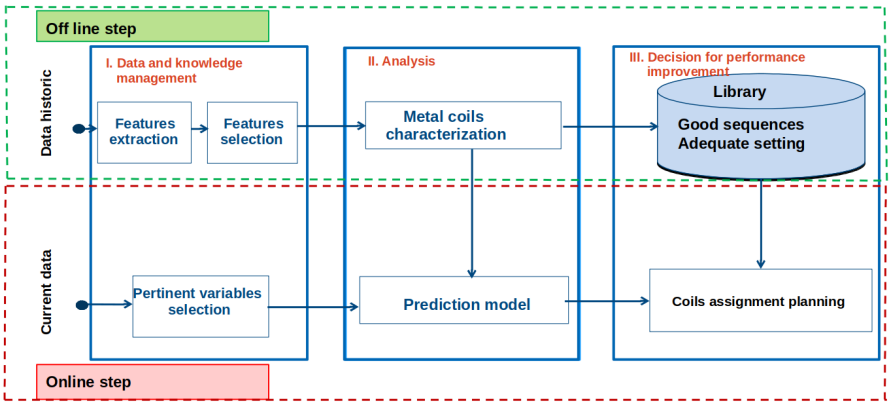


FIGURE 5.2: The proposed resolution approach.

In this work, we adopt the second possibility since the first does not represent the problem reality which will strongly impact the results. Figure 5.2 shows the adopted solution principle. The first step is to develop a model to predict each coil's performance based on its mechanical and chemical characteristics and the state of health of the tool. Then, a CBR approach is used to identify the best material allocation, ensuring stable production and an acceptable parts quality level.

The history of coils consumption for each tool and its breakdowns is saved in a referenced database used to identify the correct production sequence and the proper adjustment. When a new production is launched, the die health state is assessed as well as the available coils stock. A CBR-based algorithm estimates the impact of each coil of material based on its order of passage and its mechanical and chemical characteristics.

Thus, the best coil sequence to assign to the die is identified. In addition, for each coil, an adequate setting is also identified to optimize production.

The proposed resolution approach is encapsulated in SCODER Data System (DS2) software, which is presented in the next section.

5.3 DS2 presentation through the SCODER case study

The developments detailed previously remain theoretical models which are not applicable to their current states in the industrial domain. Indeed, these works require an enormous effort to make them exploitable. This issue concerns the deployment phase of a PHM approach. For this reason, we propose here to formalize this thesis work in a single software to guarantee its ease of use. The DS2 software is developed to encapsulate the various works carried out in this thesis. Thus, DS2 is organized in four layers (see Figure 5.3).

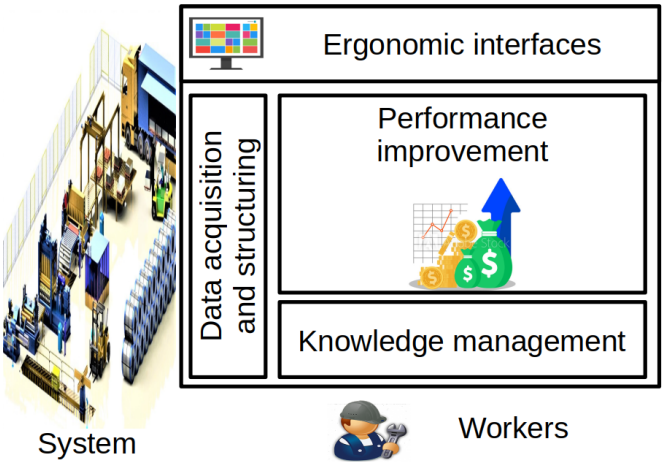


FIGURE 5.3: DS2 application architecture.

The first layer is dedicated to data acquisition and structuring. Thus, data quality metrics, and their improvement techniques presented in Chapter 2 are used in this layer. Layer 2 concerns the knowledge management loop presented in Chapter 3. Layer 3 relies on the first two layers to improve performance. Thus, this layer encapsulates in part the work developed in Chapter 4 while proposing decision support solutions. The last layer of the DS2 application concerns the presentation of these functionalities in ergonomic and interactive interfaces. In the following paragraphs, we present in detail the different layers of the DS2 software. However and regarding the multitude of existing software solutions, it is judicious to position the DS2 software among these solutions.

5.3.1 Discussion on the DS2 positioning

Figure 5.4 represents the various data technologies in an industrial enterprise. From top to bottom, this pyramid is broken down into enterprise resource planning (ERP), manufacturing execution systems (MES), Control Level, and Device Level.

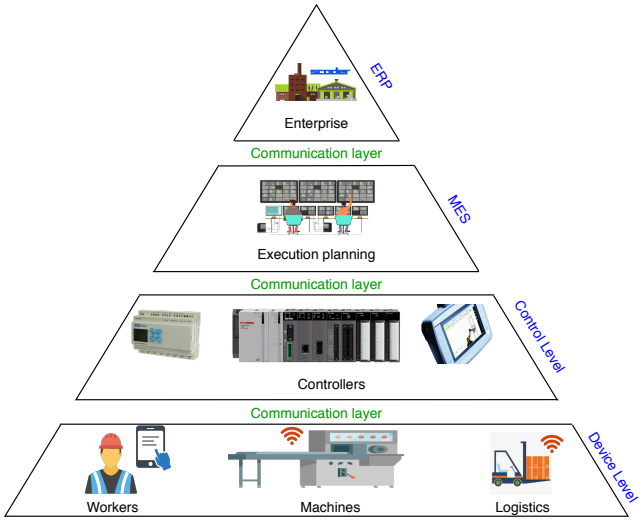


FIGURE 5.4: Automation pyramid in the industrial domain.

Recall that ERP systems are defined as a method of planning and controlling the resources required to fulfill customer orders [Ross and Vitale, 2000]. The ERP database includes information on sales, historical production data, accounting, and production line [Madanhire and Mbohwa, 2016]. However, ERP systems are accused of being adequate for big corporations where the decision process is decentralized, which is not the SMEs case [Moeuf, 2018]. Moreover, ERP systems don't allow real-time control of the production process [Kletti, 2007].

Unlike ERP systems, MES software focuses on digitizing the production process to enable real-time control of the various activities [Coronado et al., 2018]. Using real-time data, an MES system guides initiates, intervenes, and reports on workshop activities as they occur [Saenz de Ugarte et al., 2009]. This data relates to manufacturing instructions, design engineering data, the state of resources, the progress of activities, and all events during production activities.

It seems like MES can integrate the developed works in this thesis. However, this solution is very holistic and combines many functionalities that are not adequate with SMEs, generating unnecessary costs. The most crucial point is the data quality issue. MES focuses on performance improvement with sophisticated algorithms without considering the quality of the used data as input. In this context, data quality is considered as one of the most critical issues that limit the digital transformation within SMEs [Omri et al., 2021]. For that, the DS2 software is developed to fill this gap and propose a

scalable tool that fits the SMEs' needs with a moderate cost. Thus, DS2 does not replace existing technologies, but rather it is a framework to manage better the data flow generated by these various technologies.

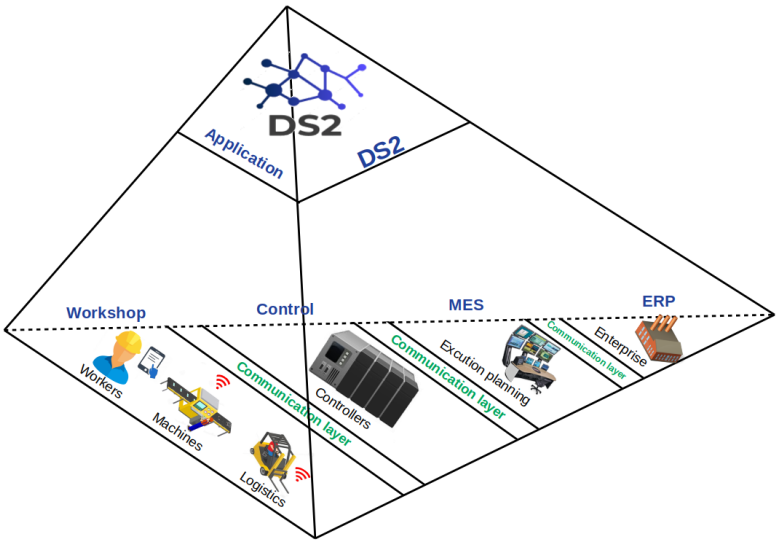


FIGURE 5.5: Work positioning regarding existing solutions.

As shown in figure 5.5, the DS2 software interacts with the various data technologies existing in the company to collect the data necessary for each application. Indeed, DS2 collects logistics data such as production orders. The data concerning the machines' operation are retrieved directly from a network of sensors installed on the machine. To ensure maximum use of the knowledge of operators, the latter participate in this approach by introducing some data concerning the state of the production process. These data are entered via DS2 interfaces installed on a network of tablets and computers throughout the workshop. Thus, DS2 ensures efficient management of the data circulating between the various teams.

5.3.2 DS2 functionalities

Figure 5.6 presents the different functionalities of the DS2 software. Indeed, this software allows to:

- **Data structuring:** It refers to the data collection and their structuring into databases used to train artificial intelligence (AI) models.
- **Equipment health assessment:** This functionality concerns the assessment of the health of machines or production systems. This phase is crucial in the proposed approach because it identifies anomalies, assesses the system health state with appropriate metrics, and guides the data analysis phase.

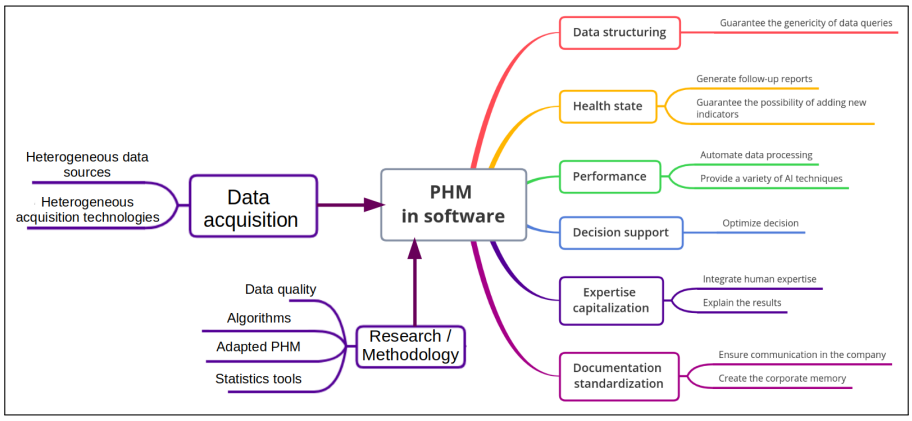


FIGURE 5.6: Presentation of the DS2 functionalities.

- **Performance prediction:** Most of our work has been incorporated into this functionality. Indeed, it allows to classify the metal coils stock, predict the failure of systems, and offer a wide range of AI tools to meet the needs of staff.
- **Decision support:** All the conducted work to optimize the metal coils assignment is integrated into this functionality.
- **Expertise capitalization:** This functionality stems from a real industrial need. Indeed, companies and, in particular, SMEs are struggling to capitalize and share their know-how. Thus, the DS2 application allows capitalizing the operator's know-how, capitalizing the extracted knowledge from the data, validating it, and efficiently sharing it throughout the company. These various functions are ensured through a collaborative PHM approach focused on workers and promoting their activities.
- **Documentation standardization:** This last feature represents the fruit of this work. It allows creating a corporate memory and, therefore, the transfer of know-how between the various teams and generations.

To guarantee the ease of use of the software, these various functionalities have been programmed in ergonomic interfaces. Recall that these interfaces are developed to be used in a french company, for that all of them are in the french language. In the following paragraphs, we detail these interfaces following the DS2 architecture.

5.3.3 Data management interfaces

This module is dedicated to management. For that, several interfaces for data collection, structuring, and improving their quality are developed based on the results of Chapter 1 and Chapter 2.

Data structuring

As shown in Figure 5.7, this interface allows launching queries to collect data from the various data sources (i.e., sensors, ERP, etc.). These data can be raw or not which characterize its management mode. Then, these data are cross-referenced to describe each consumed metal coil.

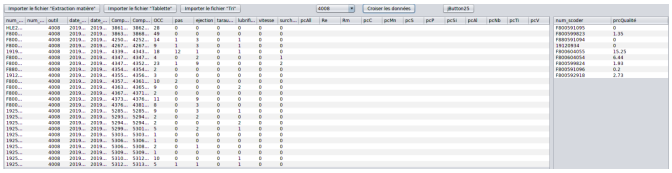


FIGURE 5.7: Presentation of the data structuring interface.

The created databases quality will be assessed and preprocessed to improve their quality and then used for training AI models and classifying the coil stock.

Data quality assessment

Before analyzing the collected data, a crucial step must be conducted, which the data quality improvement. To efficiently accomplish this task, DQ should be assessed at first. Based on the DQ metrics proposed in Chapter 2, a dedicated interface is developed to quantify DQ regarding the most encountered DQ issues (see Figure 5.8).

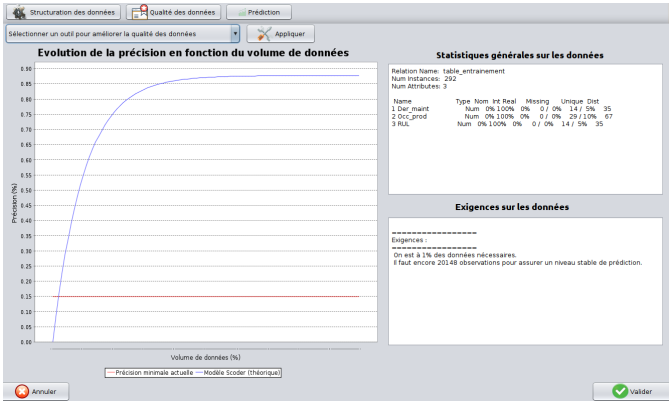


FIGURE 5.8: Presentation of the DQ assessment interface.

Moreover, this interface is used to identify DQ requirements. Based on the empirical models developed in Chapter 4, DQ requirements are identified regarding a fixed performance level. The objective is to optimize DQ by reducing the acquisition cost.

Based on the identified DQ requirements, several data quality improvement techniques can be used to meet these requirements. The dedicated DQ improvement interface is presented in the following paragraph.

Data quality improvement

This interface deals with the data quality improvement problem. Thus, the user can use it to overcome shortness present in databases (i.e., missing data, noisy data, etc.).

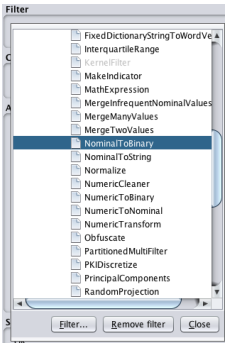


FIGURE 5.9: Data quality improvement interface.

As shown in Figure 5.9, this interface proposes an extensive list of traditional data quality improvement techniques detailed in Chapter 2. Thus, the user can choose from this list to deal with data quality issues. One should note that the developed knowledge-based approach for data quality improvement is not integrated into this interface. However, it is integrated into the knowledge management module, which we detail in the following paragraph.

5.3.4 Knowledge management interfaces

As mentioned throughout this manuscript, the user is at the center of the proposed approach. Indeed, the operator has an idea of the problem, its causes, and its consequences. He generally links each problem to a set of variables that can affect it (or the variable with no influence). Thus, this knowledge will be formalized in suitable rules. This set of rules are then used to perform the data quality improvement. As detailed in Chapter 3, there are several ways to integrate human knowledge. The operators' expertise is used to improve the data quality in four ways:

- Generate new variables.
- Identify the features importance.
- Describe statistically each variable.
- Identify statistical relationships between variables.

Figure 5.10 shows the interface which allows the user to choose between this options to integrate his knowledge. In addition to these knowledge management methods, operators can communicate information about the machine through free comments. These

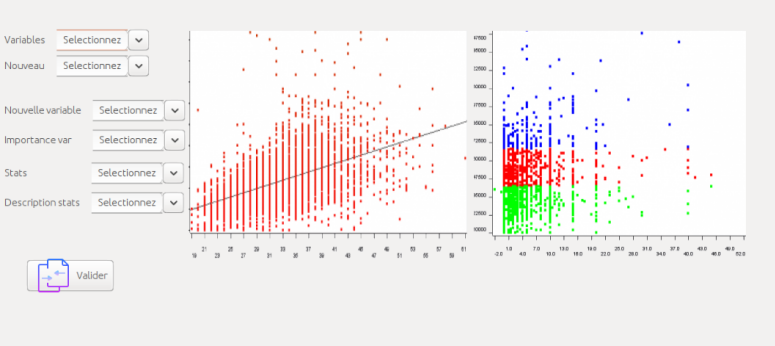


FIGURE 5.10: Knowledge management interface.

comments are mainly used to facilitate communication between the various teams. Also, they feed a natural language processing (NLP) algorithm to extract knowledge which needs more data to be globally validated.

As detailed in Chapter 3, the second phase of the knowledge management module explains the results. A dedicated interface is developed to explain the data analysis results (see Figure 5.11).

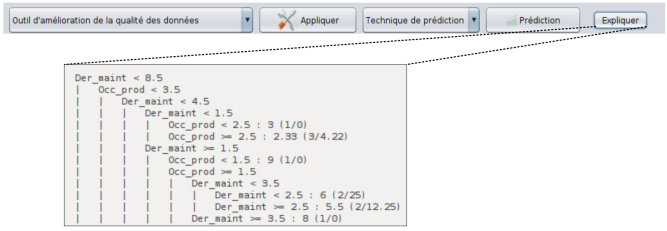


FIGURE 5.11: Results explanation interface.

Thanks to the DS2 application, the user can choose between two explanation solutions. The first one provides a semi-global explanation as detailed in the Chapter 3 while the second offers a global explanation using a decision tree (DT) model.

5.3.5 Performance improvement interfaces

After the data acquisition and the improvement of their quality, data are used to improve the system performance. This phase is conducted in three steps loop: Assessment - Analysis - Decision.

Health state assessment via KPIs

The first step of this process is to assess the equipment operation through a set of key performance indicators (KPIs). These KPIs are the same used within the SCODER

company. Thanks to the DS2 application, these indicators are calculated and displayed automatically. Figure 5.12 shows the dedicated interface to display maintenance indicators. It is mainly used by the maintenance team to assess machines and prepare interventions. Also, it helps production team to make reliable production plannings.

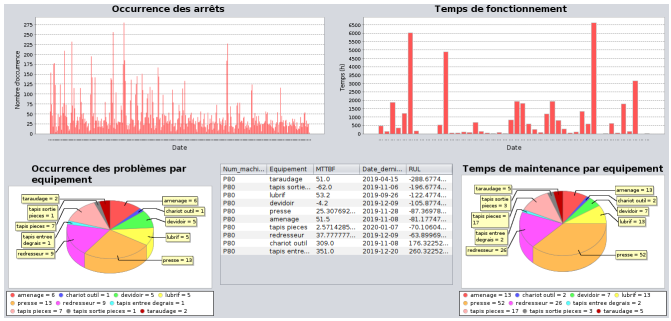
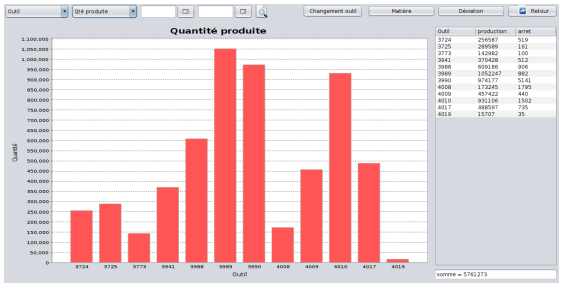
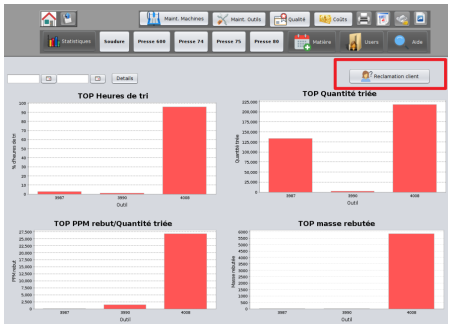


FIGURE 5.12: Maintenance indicator computation and visualization.

In addition to this interface, two others are developed to compute production and parts quality indicators automatically. The first interface (see Figure 5.13a) calculates indicators on the machine productivity such as downtime occurrences, operating time, produced parts number, etc. While the second interface (see Figure 5.13b) presents statistics on sorted parts such as percentage of sorted parts by die, the mass and rate of rejected parts, and the inspection time.



(a) Productivity indicators interface.



(b) Parts quality indicators.

FIGURE 5.13: DS2's interfaces presentation.

This automation of the KPIs computing allows to report the equipment health state and thus quickly identify anomalies. Once the anomalies are identified, the analysis and resolution phase begins. The following paragraph details this phase.

Problem analysis

The problem analysis module aims to make the data analysis phase easier to ensure an informed decision. Thus, several AI tools have been developed to facilitate the data analysis tasks, such as feature selection, clustering, and prediction (see Figure 5.14).

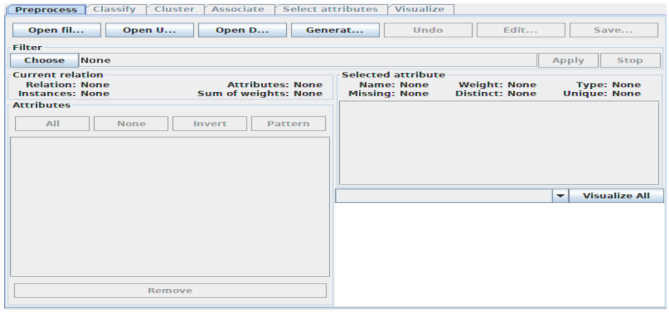
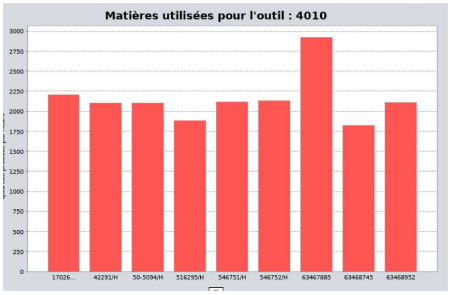
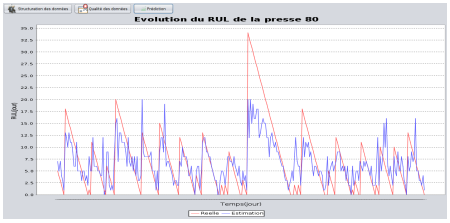


FIGURE 5.14: Presentation of the AI tools interface.

These different techniques are used to cross input variables to predict an output (or several outputs). As shown in Figure 5.15, two main applications are conducted: predict each metal coil's impact on the production according to its characteristics and predict the machine remaining useful life (RUL).



(a) Metal coil's impact prediction.



(b) RUL prediction.

FIGURE 5.15: DS2's interfaces presentation.

Based on the information resulting from these predictions, a decision module is developed to optimize the production performance. The following paragraph details this module.

Decision support

We come here to the decision support, which depends on all the previously detailed interfaces' functionalities. Data are collected, structured, their quality assessed and improved, and used to detect anomalies and predict their occurrence. These predictions are

then used to feed a decision support module for performance improvement. This module's main application consists of a new strategy for metal coils assignment optimization to improve performance in terms of productivity and part quality. To do this, a CBR-based algorithm is integrated into the DS2 software to identify the best coil sequences and adequate dies configurations. This task is performed while respecting a set of constraints:

- No bad metal coils can be affected overnight.
- Planning can only be done for an horizon in which the coil stock is known.

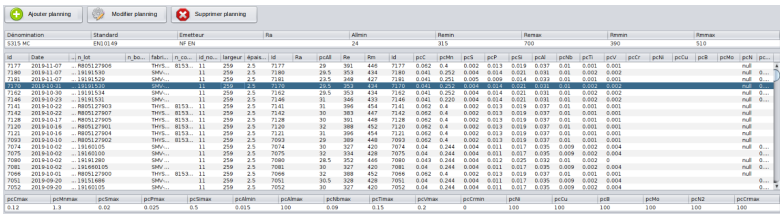
While respecting these constraints, a dataset was built using DS2 to solve this problem. This dataset consists of a Metal coil Identity Card (MIC) where each coil's characteristics, its result in the production process, and the corresponding die configurations are saved. The characteristics of this database are presented in Table 5.2.

TABLE 5.1: Presentation of the coils assignment interface.

Dataset	**# instances**	**# variables**	**# classes**
MIC	90	24	2

TABLE 5.2: Characteristics of the MIC dataset.

The resolution approach proposed in Section 5.2.2 is used to analyze this dataset and present a coils assignment plan. Figure **??** shows the dedicated interface to accomplish this task automatically.



Recall that the metal coils assignment application is only used to clarify and describe the different interfaces of the DS2 software. However, many other applications can be implemented in the DS2 software for performance optimization, such as workforce planning, maintenance actions planning, machines' operation optimization, etc. The DS2 software has been successfully installed within the SCODER company, where the obtained results are detailed in the next section.

5.4 SCODER case study global validation

In this section, we present the results of the application of this work within the SCODER company. Despite the novelty of the PHM project, the performance of the studied machine has been improved. The figure 5.16 shows the evolution of the produced parts quantity between two breakdowns during 115 days of production. The PHM study began on day 50 of production. This indicator allows quantifying the occurrence of breakdowns taking into account the produced quantity. The results show that productivity was improved by over 80% after the deployment of PHM.

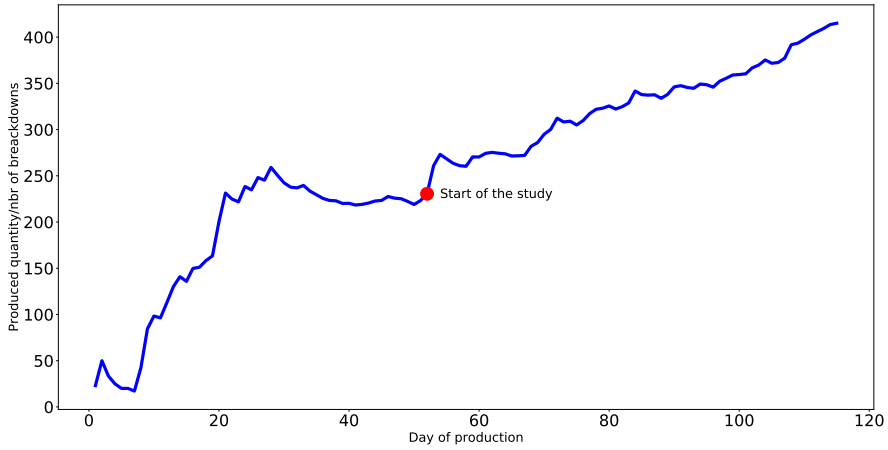


FIGURE 5.16: Study impact on the machine productivity.

Along with this productivity improvement comes an improvement in part quality, from a scrap rate of 2 % to a rate close to 0 %. Additionally, the machine's TRS increased from 65 % to 79 %.

Remember that this thesis's scientific challenge is to make complex concepts such as Big Data, the Connected Object, artificial intelligence accessible for an SME like SCODER to strengthen its competitiveness and put it on the rails of Industry 4.0. The challenge has been met, and the results are satisfactory. On the other hand, this work must be maintained and improved to achieve the long-term objectives of SCODER. There is also a real human stake, convincing staff that this project is not to expel them but rather to help them make their activities more efficient. Beyond the SCODER company, this work creates new opportunities to strengthen the region's industrial fabric and participate modestly in improving their contribution to the national economy.

5.5 Conclusion

The objective of this chapter is two-folded: (1) a software implementation of the proposed approach, and (2) a synthesis of the global validation through SCODER case study. For

that purpose, the SCODER Data System (DS2) is presented as an encapsulation of the different phases of this thesis work. Then, the validation process is illustrated through the SCODER case study for metal coils assignment optimization. Three steps are considered: data management, knowledge management, and performance improvement. The obtained results meet the industrial expectations.

The conducted work is partially validated and needs to be enriched with more significant case studies.

General Conclusion and Perspectives

"Difficult roads often lead to beautiful destinations."

-anyone

General conclusion

Data technologies' emergence brings many developments in the industrial domain and pushed companies on the road to digitization. Nevertheless, Industry 4.0, Big data, and artificial intelligence are concepts more associated with large companies. One of the current challenges for SMEs is a reflection on these concepts' appropriation to avoid a competitive gap with large groups. In this context, the SCODER company is leading a voluntarist digitization strategy. One of this strategy's axes corresponds to scientific collaboration with the Femto-st institute and the ENSMM of Besançon (CIFRE thesis). The objective was to develop a methodology for Industry 4.0 integration within SMEs. A bibliographical study reinforced by field observations was conducted to identify the parameters that allow this objective's satisfaction. These parameters can be summarized in three points: data structuring, knowledge management, and performance improvement at the right cost. Indeed, the advantage in SMEs is that the operators are close to the process, which allows them to identify their needs and prioritize them in terms of urgency. The staff's versatility allows them to analyze each problem concerning its impact on the entire production process. Thus, the problem modeling phase and identifying the required variables to solve the problem seem more straightforward and more efficient in SMEs, which reduces implementation costs.

The objective of this work is to facilitate digital transformation in SMEs. To do so, the PHM has been used as a framework for (i) data digitization, (ii) knowledge management, and (iii) industrial performance improvement at the right cost. For this reason, a large part of this work has been devoted to develop these points and to propose adequate solutions and thus successfully integrate Industry 4.0 within SMEs.

The **Chapter 1** allows the positioning of this thesis work through the SCODER industrial problem statement. The industrial problem statement emerges from a concrete use case, materialized by a stamping line in the SCODER factory. It addresses the challenge of implementing industry 4.0 technologies within SMEs. Thus, a bibliographic study was conducted to identify the barriers that limit the Industry 4.0 implementation within SMEs. Based on these barriers, an adapted PHM approach is proposed as a solution to implement Industry 4.0 within SMEs. This approach is applied in the SCODER company, which permits identifying the data quality management issue as the most critical research gaps. Thus it is considered as the backbone of this thesis work.

In **Chapter 2**, the literature on data quality management is extensively reviewed. The most encountered data quality issues in the PHM context (i.e., imbalanced, missing, and noisy data) are studied upon different criteria related to two essential domains: 1) quantification metrics and 2) improvement techniques. The study shows that data quality improvement is a serious problem in the SME context. Technological solutions to improve data quality or collect new variables are not considered because they can generate significant costs unsupportable by SMEs. Concerning the software improvement techniques, they are not applicable when variables are not recorded or are highly damaged, which is the case of SMEs. Faced with these problems, traditional data quality improvement solutions are not suitable for SMEs. Thus the quality improvement of highly damaged variables is identified as the main research gap.

Whatever the data analysis technique, it requires relevant data that satisfy the required data quality for the fixed objective. However, data in SMEs are generally highly damaged, which limits their improvement by traditional techniques. To deal with this problem, the **Chapter 3** proposes a new knowledge-based methodology for data quality improvement. Thus, a generic formalization of the operators' know-how is introduced to improve data quality and improve the data analysis results. For this purpose, human knowledge types are defined, and their integration modes in the data management process are discussed. Based on this study, a knowledge-based data quality improvement framework is proposed to deal with the data quality improvement issues in the SME context. In a second step and to ensure efficient knowledge management, the obtained analysis results are explained and used to enrich the human knowledge applied to improve data quality in future analysis tasks. The proposed approach is applied in the SCODER case study in order to validate it and assess its applicability. The obtained results allow to capitalize the expertise and contribute in part to the corporate memory creation.

To optimize the data quality costs and improve performance, the **Chapter 4** presents a novel methodology for performance improvement at the right cost. To do, the data quality impact on the fault detection task of the PHM process is formalized. This formalization is used to propose an empiric metric to quantify this impact. Based on this empirical metric, a methodology to improve industrial performance at the right cost has been proposed. However, imposing a fixed cost for a data project may block its imple-

mentation even if this cost is optimized. To overcome this problem, another methodology was presented to maximize performance with a fixed budget. This work was applied in the SCODER company to optimize the data acquisition infrastructure. The obtained results are encouraging and prove the applicability of this work in the industrial domain.

The developed approaches in this thesis are validated as things progress at the end of each chapter. Moreover, a global validation is proposed in the **Chapter 5**. The developed methodology is applied in the SCODER company, an SME installed near Besançon and specialized in high precision cutting for automotive applications. To facilitate its implementation, all the steps of this methodology have been encapsulated in SCODER Data System (DS2) software. Thus, this chapter has been devoted to present this software and its functionalities through a real case study of metal coils assignment optimization. The installation steps are described, and the obtained results are reported.

To sum up, five main contributions were elaborated in this thesis:

- Data quality issues formalization for industry 4.0 within SMEs.
- Data quality management based on Knowledge oriented methodology.
- Impact of the data quality on industrial performance.
- Data quality optimization for performance improvement at the right cost.
- Industrial contribution: DS2 software implementation to perform all above contributions.

Given the innovative and exploratory nature of the thesis topic, the proposed methodology shows good potential and meets the industrialist's objectives. The obtained results open many perspectives, which are presented in the next paragraph.

Perspectives

The multidisciplinary nature of this thesis work allowed, on the one hand, to cover a wide range of themes and to propose, on the other hand, several perspectives. This paragraph proposes to classify these perspectives according to the particularly concerned theme: data quality management, knowledge management, cost optimization, and human and social sciences. Future directions of this work are ordered in descending order according to their complexity level. The lowest level means an optimization of a developed method or its adaptation with other hypotheses. The higher levels represent a more substantial experimental, technological and/or methodological development.

Data quality management

Data quality management is a crucial topic in the digitization domain. The ideal in this context is to carry out a complete theoretical study to quantify the impact of data

quality on the PHM. Such a study does not seem easy and requires skills in statistics and probability rather than data analysis. In the absence of a complete analytical formulation, an empirical analysis may help understand the impact of data qualities on a PHM process's different tasks. Before this, it would be judicious to complete data quality analysis in the PHM context by proposing adequate metrics to quantify them on the theoretical level.

Knowledge management

We have seen, through this work, that the knowledge aspect opens up immense possibilities for improving industrial performance. Nevertheless, a greater mastery of human knowledge is more than required. Thus, several perspectives can be considered in this context, such as the proposal of a complete formalization of this knowledge. This formalization can be useful in improving data quality and the rest of the data analysis process. However, human knowledge can contain bias which strongly affects the analysis results. Thus, further works should be conducted for knowledge uncertainties quantification and assessment of their impact on the entire data analysis process. Also, one should think about the development of new approaches for the correction of these uncertainties.

Data quality cost optimization

The main difference between MNEs and SMEs resides in resource availability. For that, each digitization strategy dedicated to SMEs should take into account this aspect. In this context, perspectives can go into developing an analytical data quality cost model. Based on this cost modeling, the economic part of digitization strategies can be addressed efficiently. Such a model allows to identify the required investment and quantifying the expected profits.

Human and social sciences

Beyond these engineering sciences perspectives, it remains a real human and social issue which is the human dimension. This dimension should be supported and prepared to improve the new industrial ecosystem (digitization and knowledge transmission).

Given the innovative and exploratory nature of the thesis topic, the proposed perspectives maybe not exhaustive. Only **more real case studies** can reveal the real research gaps that should be filled.

Bibliography

[DAM, 2017] (2017). Dama-dmbok (2nd edition): Data management body of knowledge. *DAMA International - 2017*, pages i–. 39

[Adams et al., 2017] Adams, S., Malinowski, M., Heddy, G., Choo, B., and Beling, P. A. (2017). The wear methodology for prognostics and health management implementation in manufacturing. *Journal of Manufacturing Systems*, 45:82–96. 22

[Amin and Ali, 2018] Amin, M. and Ali, A. (2018). Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decisions. *Wavy AI Research Foundation: Lahore, Pakistan*. 90

[Antunes et al., 2008] Antunes, P., Herskovic, V., Ochoa, S. F., and Pino, J. A. (2008). Structuring dimensions for collaborative systems evaluation. *ACM computing surveys (CSUR)*, 44(2):1–28. 66

[Arteaga and Ferrer, 2005] Arteaga, F. and Ferrer, A. (2005). Framework for regression-based missing data imputation methods in on-line mspc. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(8):439–447. 52

[Babiceanu and Seker, 2016] Babiceanu, R. F. and Seker, R. (2016). Big data and virtualization for manufacturing cyber-physical systems: A survey of the current status and future outlook. *Computers in Industry*, 81:128–137. 16

[Bai et al., 2016] Bai, M., Wang, X., Xin, J., and Wang, G. (2016). An efficient algorithm for distributed density-based outlier detection on big data. *Neurocomputing*, 181:19–28. 54

[Batini et al., 2009] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):16. 45, 47

[Batini and Scannapieco, 2016] Batini, C. and Scannapieco, M. (2016). Data and information quality: Concepts, methodologies and techniques. 48

[Beaulieu-Jones et al., 2017] Beaulieu-Jones, B. K., Moore, J. H., and CONSORTIUM, P. R. O.-A. A. C. T. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. In *Pacific Symposium on Biocomputing 2017*, pages 207–218. World Scientific. 52

[Beyan and Fisher, 2015] Beyan, C. and Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, 48(5):1653–1672. 50

[Bidet-Mayer and Ciet, 2016] Bidet-Mayer, T. and Ciet, N. (2016). *L'industrie du futur: une compétition mondiale*. La Fabrique de l'Industrie. 17

[Bifet et al., 2010] Bifet, A., Holmes, G., Kirkby, R., and Pfahringer, B. (2010). Moa: Massive online analysis. *Journal of Machine Learning Research*, 11(May):1601–1604. 54

[Blind and Mangelsdorf, 2012] Blind, K. and Mangelsdorf, A. (2012). Alliance formation of smes: empirical evidence from standardization committees. *IEEE Transactions on Engineering Management*, 60(1):148–156. 15

[Boden et al., 2012] Boden, A., Avram, G., Bannon, L., and Wulf, V. (2012). Knowledge sharing practices and the impact of cultural factors: reflections on two case studies of offshoring in sme. *Journal of software: Evolution and Process*, 24(2):139–152. 15

[Bohanec and Rajkovic, 1988] Bohanec, M. and Rajkovic, V. (1988). Knowledge acquisition and explanation for multi-attribute decision making. In *8th Intl Workshop on Expert Systems and their Applications*, pages 59–78. 86

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. 74, 85

[Breunig et al., 2000] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104. 54

[Bridge and O'Neill, 2012] Bridge, S. and O'Neill, K. (2012). *Understanding enterprise: Entrepreneurship and small business*. Macmillan International Higher Education. 20

[Brown et al., 1998] Brown, A., Van Der Wiele, T., and Loughton, K. (1998). Smaller enterprises' experiences with iso 9000. *International journal of quality & reliability management*. 15

[Bublitz and Noseleit, 2014] Bublitz, E. and Noseleit, F. (2014). The skill balancing act: when does broad expertise pay off? *Small Business Economics*, 42(1):17–32. 14, 21

[Caballero et al., 2014] Caballero, I., Serrano, M., and Piattini, M. (2014). A data quality in use model for big data. In *International Conference on Conceptual Modeling*, pages 65–74. Springer. 46

[Cai and Zhu, 2015] Cai, L. and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14. 46

[Campos et al., 2016] Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*, 30(4):891–927. 52, 53

[Canada, 2021] Canada, G. (2021). Sme research and statistics. *http://www.ic.gc.ca/eic/site/061.nsf/eng/Home*, page (accessed: 22.01.2021). 12

[Cappiello et al., 2004] Cappiello, C., Francalanci, C., and Pernici, B. (2004). Data quality assessment from the user's perspective. In *Proceedings of the 2004 international workshop on Information quality in information systems*, pages 68–73. 46

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357. 50

[Chen et al., 2013] Chen, Y., Zhu, F., and Lee, J. (2013). Data quality evaluation and improvement for prognostic modeling using visual assessment based data partitioning method. *Computers in industry*, 64(3):214–225. 47

[Cheng et al., 2010] Cheng, S., Azarian, M. H., and Pecht, M. G. (2010). Sensor systems for prognostics and health management. *Sensors*, 10(6):5774–5797. 28

[Chhabra et al., 2017] Chhabra, G., Vashisht, V., and Ranjan, J. (2017). A comparison of multiple imputation methods for data with missing values. *Indian Journal of Science and Technology*, 10(19):1–7. 52

[China.org.cn, 2021] China.org.cn (2021). Law of the people's republic of china on promotion of smes. *http://www.china.org.cn/english/government/207325.htm*, page (accessed: 22.01.2021). 12

[Cooper et al., 2004] Cooper, W. W., Seiford, L. M., and Zhu, J. (2004). Data envelopment analysis. In *Handbook on data envelopment analysis*, pages 1–39. Springer. 29

[Coronado et al., 2018] Coronado, P. D. U., Lynn, R., Louhichi, W., Parto, M., Wescoat, E., and Kurfess, T. (2018). Part data integration in the shop floor digital twin: Mobile and cloud technologies to enable a manufacturing execution system. *Journal of manufacturing systems*, 48:25–33. 41, 104

[Cui and Xia, 2017] Cui, D. and Xia, K. (2017). Strip surface defects recognition based on pso-rs&socp-svm algorithm. *Mathematical Problems in Engineering*, 2017. 50

[Czerniak and Zarzycki, 2003] Czerniak, J. and Zarzycki, H. (2003). Application of rough sets in the presumptive diagnosis of urinary system diseases. In *Artificial intelligence and security in computing systems*, pages 41–51. Springer. 90

[Dang et al., 2015] Dang, T. T., Ngan, H. Y., and Liu, W. (2015). Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 507–510. IEEE. 54

[Dangayach and Deshmukh, 2005] Dangayach, G. and Deshmukh, S. (2005). Advanced manufacturing technology implementation: evidence from indian small and medium enterprises (smes). *Journal of Manufacturing Technology Management*. 14

[Das et al., 2008] Das, K., Schneider, J., and Neill, D. B. (2008). Anomaly pattern detection in categorical datasets. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–176. 54

[Das et al., 2016] Das, S., Wong, W.-K., Dietterich, T., Fern, A., and Emmott, A. (2016). Incorporating expert feedback into active anomaly discovery. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 853–858. IEEE. 54

[Datta and Das, 2015] Datta, S. and Das, S. (2015). Near-bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs. *Neural Networks*, 70:39–52. 50

[Deng, 2021] Deng, W. (2021). Industry 4.0: Digitization, sensorization and optimization. *https://inform.tmforum.org/features-and-analysis/2016/10/industry-4-0-digitization-sensorization-optimization/*, page (accessed: 26.01.2021). 17

[Dijkman et al., 2015] Dijkman, R. M., Sprenkels, B., Peeters, T., and Janssen, A. (2015). Business models for the internet of things. *International Journal of Information Management*, 35(6):672–678. 16

[Ding, 2008] Ding, S. X. (2008). *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer Science & Business Media. 83

[Dombrowski et al., 2010] Dombrowski, U., Crespo, I., and Zahn, T. (2010). Adaptive configuration of a lean production system in small and medium-sized enterprises. *Production Engineering*, 4(4):341–348. 14, 21

[Donders et al., 2006] Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091. 52

[Dong and Srivastava, 2015] Dong, X. L. and Srivastava, D. (2015). Big data integration. *Synthesis Lectures on Data Management*, 7(1):1–198. 46

[Du et al., 2015] Du, H. et al. (2015). Robust local outlier detection. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 116–123. IEEE. 54

[Durst and Bruns, 2018] Durst, S. and Bruns, G. (2018). Knowledge management in small and medium-sized enterprises. In *The Palgrave Handbook of Knowledge Management*, pages 495–514. Springer. 15

[Durst and Runar Edvardsson, 2012] Durst, S. and Runar Edvardsson, I. (2012). Knowledge management in smes: a literature review. *Journal of Knowledge Management*, 16(6):879–903. 20

[Eppler and Helfert, 2004] Eppler, M. and Helfert, M. (2004). A classification and analysis of data quality costs. In *International Conference on Information Quality*, pages 311–325. 81

[Erfani et al., 2016] Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134. 54

[Feldman et al., 2008] Feldman, K., Sandborn, P., and Jazouli, T. (2008). The analysis of return on investment for phm applied to electronic systems. In *2008 International Conference on Prognostics and Health Management*, pages 1–9. IEEE. 25

[Filannino, 2011] Filannino, M. (2011). Dbworld e-mail classification using a very small corpus. *The University of Manchester*. 86

[Fischer, 1992] Fischer, A. (1992). A special newton-type optimization method. *Optimization*, 24(3-4):269–284. 95

[Fisher, 2009] Fisher, T. (2009). *The data asset: How smart companies govern their data for business success*, volume 24. John Wiley & Sons. 39

[France-industrie, 2020] France-industrie (2020). L'industrie en france c'est : Tous les chiffres clés. *https://www.franceindustrie.org/*, page (accessed: 04.12.2020). 3, 12

[Freitas, 2014] Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10. 66

[Fung et al., 2005] Fung, G., Sandilya, S., and Rao, R. B. (2005). Rule extraction from linear support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 32–40. 65

[Galar et al., 2011] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484. 49, 50

[Gao et al., 2018] Gao, X., Niculita, O., Alkali, B., and McGlinchey, D. (2018). Cost benefit analysis of applying phm for subsea applications. In *Proceedings of the European Conference of the PHM Society*. 25

[Goldstein and Dengel, 2012] Goldstein, M. and Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63. 54

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680. 72

[Graham, 2003] Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling*, 10(1):80–100. 52

[Guidotti et al., 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42. 59, 65, 66

[Guillén et al., 2016] Guillén, A. J., González-Prida, V., Gómez, J. F., and Crespo, A. (2016). Standards as reference to build a phm-based solution. In *Proceedings of the 10th World Congress on Engineering Asset Management (WCEAM 2015)*, pages 207–214. Springer. 23, 26, 41

[Guo et al., 2019] Guo, Y., Wang, L., Wang, M., and Zhang, X. (2019). The mediating role of environmental innovation on knowledge acquisition and corporate performance relationship—a study of smes in china. *Sustainability*, 11(8):2315. 12

[Haug et al., 2011] Haug, A., Zachariassen, F., and Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)*, 4(2):168–193. 81

[He and Garcia, 2009] He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284. 67

[He et al., 2003] He, Z., Xu, X., and Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650. 54

[Helu et al., 2015] Helu, M., Morris, K., Jung, K., Lyons, K., and Leong, S. (2015). Identifying performance assurance challenges for smart manufacturing. *Manufacturing letters*, 6:1–4. 14

[Hido et al., 2009] Hido, S., Kashima, H., and Takahashi, Y. (2009). Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2(5-6):412–426. 50

[Hodge and Austin, 2004] Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126. 48

[Hoffmann et al., 2016] Hoffmann, M., Büscher, C., Meisen, T., and Jeschke, S. (2016). Continuous integration of field level production data into top-level information systems using the opc interface standard. *Procedia CIRP*, 41:496–501. 41

[Hooda et al., 2018] Hooda, N., Bawa, S., and Rana, P. S. (2018). Fraudulent firm classification: a case study of an external audit. *Applied Artificial Intelligence*, 32(1):48–64. 86

[Institute, 2021] Institute, D. G. (2021). Data governance institute framework. *https://datagovernance.com/*, page (accessed: 04.02.2021). ix, 44

[Islam, 2013] Islam, M. S. (2013). An assessment for focusing the change of data quality (dq) with timeliness in information manufacturing systems. In *Proceedings of the SDIWC Second International Conference on Digital Enterprise and Information Systems (DEIS2013), Kuala Lumpur, Malaysia*, pages 4–6. 46

[ISO, 2003] ISO (2003). *SO 13374-1:2003 - Condition monitoring and diagnostics of machines Data processing, communication and presentation Part 1: General guidelines.* 24

[ISO, 2007] ISO (2007). *ISO 13374-2:2007 - Condition monitoring and diagnostics of machines Data processing, communicaon and presentaon Part 2: Data processing.* 24

[ISO, 2012] ISO (2012). *ISO 13374-3:2012 - Condition monitoring and diagnostics of machines Data processing, communicaon and presentaon Part 3: Communication.* 25

[ISO, 2015] ISO (2015). *ISO 13374-4:2015 Condition monitoring and diagnostics of machine systems – Data processing, communication and presentation – Part 4: Presentation.* 25

[ISO/IEC, 2008] ISO/IEC (2008). Software engineering software product quality requirements and evaluation (square) data quality model. In *ISO/IEC, Tech. Rep. ISO/IEC 25012, 2008.* 45

[ISO/IEC, 2015] ISO/IEC (2015). Iso 80008:2015 data quality part 8: Information and data quality: Concepts and measuring. In *ISO/IEC, Tech. Rep. ISO/IEC 8000, 2015.* 44

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice-Hall, Inc. 47

[Jardine et al., 2006] Jardine, A. K., Lin, D., and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510. 23

[Jasra et al., 2011] Jasra, J., Hunjra, A. I., Rehman, A. U., Azam, R. I., and Khan, M. A. (2011). Determinants of business success of small and medium enterprises. *International Journal of Business and Social Science*, 2(20). 14

[Jia et al., 2017] Jia, X., Zhao, M., Di, Y., Yang, Q., and Lee, J. (2017). Assessment of data suitability for machine prognosis using maximum mean discrepancy. *IEEE transactions on industrial electronics*, 65(7):5872–5881. 47, 82

[Jin et al., 2006] Jin, W., Tung, A. K., Han, J., and Wang, W. (2006). Ranking outliers using symmetric neighborhood relationship. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 577–593. Springer. 54

[Johansson et al., 2004] Johansson, U., Niklasson, L., and König, R. (2004). Accuracy vs. comprehensibility in data mining models. In *Proceedings of the seventh international conference on information fusion*, volume 1, pages 295–300. 65

[Juddoo et al., 2018] Juddoo, S., George, C., Duquenoy, P., and Windridge, D. (2018). Data governance in the health industry: Investigating data quality dimensions within a big data context. *Applied System Innovation*, 1(4):43. 44, 45, 46, 47

[Julien and Ramangalahy, 2003a] Julien, P. and Ramangalahy, C. (2003a). Competitive strategy and performance of exporting smes: An empirical investigation of the impact of their export information search and competencies. *Entrepreneurship Theory and Practice*, 27(3):227–245. 20

[Julien and Ramangalahy, 2003b] Julien, P.-A. and Ramangalahy, C. (2003b). Competitive strategy and performance of exporting smes: An empirical investigation of the impact of their export information search and competencies. *Entrepreneurship Theory and Practice*, 27(3):227–245. 15

[Jung and Jin, 2018] Jung, J.-u. and Jin, K.-h. (2018). Case studies for the establishment of the optimized smart factory with small and medium-sized enterprises. In *Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control*, pages 1–5. 19

[Kennedy et al., 2003] Kennedy, J., Hyland, P., et al. (2003). A comparison of manufacturing technology adoption in smes and large companies. In *Proceedings of 16th Annual Conference of Small Enterprise Association of Australia and New Zealand*, pages 1–10. 20, 27

[Khanzadeh et al., 2018] Khanzadeh, M., Chowdhury, S., Marufuzzaman, M., Tschopp, M. A., and Bian, L. (2018). Porosity prediction: Supervised-learning of thermal history for direct laser deposition. *Journal of manufacturing systems*, 47:69–82. 87

[Khozeimeh et al., 2017] Khozeimeh, F., Alizadehsani, R., Roshanzamir, M., Khosravi, A., Layegh, P., and Nahavandi, S. (2017). An expert system for selecting wart treatment method. *Computers in biology and medicine*, 81:167–175. 90

[King et al., 1998] King, G., Honaker, J., Joseph, A., and Scheve, K. (1998). List-wise deletion is evil: what to do about missing data in political science. In *Annual Meeting of the American Political Science Association, Boston*. 52

[Kletti, 2007] Kletti, J. (2007). *Manufacturing execution system-MES*. Springer. 104

[Kohler and Weisz, 2016] Kohler, D. and Weisz, J.-D. (2016). *Industrie 4.0: les défis de la transformation numérique du modèle industriel allemand*. La Documentation française. 17

[Kong et al., 2020] Kong, T., Hu, T., Zhou, T., and Ye, Y. (2020). Data construction method for the applications of workshop digital twin system. *Journal of Manufacturing Systems*. 39

[Kothamasu et al., 2006] Kothamasu, R., Huang, S. H., and VerDuin, W. H. (2006). System health monitoring and prognostics—a review of current paradigms and practices. *The International Journal of Advanced Manufacturing Technology*, 28(9-10):1012–1024. 21

[Koulali et al., 2018] Koulali, M.-A., Koulali, S., Tembine, H., and Kobbane, A. (2018). Industrial internet of things-based prognostic health management: a mean-field stochastic game approach. *IEEE Access*, 6:54388–54395. 21, 25

[Krawczyk et al., 2014] Krawczyk, B., Woźniak, M., and Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554–562. 50, 63

[Kusiak, 2018] Kusiak, A. (2018). Smart manufacturing. *International Journal of Production Research*, 56(1-2):508–517. 19

[Kwon et al., 2016] Kwon, D., Hodkiewicz, M. R., Fan, J., Shibutani, T., and Pecht, M. G. (2016). Iot-based prognostics and systems health management for industrial applications. *IEEE Access*, 4:3659–3670. 21, 25

[Ladickỳ et al., 2015] Ladickỳ, L., Jeong, S., Solenthaler, B., Pollefeys, M., and Gross, M. (2015). Data-driven fluid simulations using regression forests. *ACM Transactions on Graphics (TOG)*, 34(6):1–9. 62

[Lakkaraju et al., 2016] Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684. 66

[Laranjeiro et al., 2015] Laranjeiro, N., Soydemir, S. N., and Bernardino, J. (2015). A survey on data quality: classifying poor data. In *2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC)*, pages 179–188. IEEE. 46

[Leake, 1996] Leake, D. B. (1996). Case-based reasoning: experiences, lessons, and future directions. 102

[Lee et al., 2014a] Lee, J., Holgado, M., Kao, H.-A., and Macchi, M. (2014a). New thinking paradigm for maintenance innovation design. *IFAC Proceedings Volumes*, 47(3):7104–7109. 23

[Lee et al., 2014b] Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., and Siegel, D. (2014b). Prognostics and health management design for rotary machinery systems—reviews, methodology and applications. *Mechanical systems and signal processing*, 42(1-2):314–334. 47

[Lee et al., 2010] Lee, S., Park, G., Yoon, B., and Park, J. (2010). Open innovation in smes—an intermediated network model. *Research policy*, 39(2):290–300. 15

[Lei et al., 2018] Lei, Y., Li, N., Guo, L., Li, N., Yan, T., and Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to rul prediction. *Mechanical Systems and Signal Processing*, 104:799–834. 23

[Li et al., 2015] Li, J., Tao, F., Cheng, Y., and Zhao, L. (2015). Big data in product lifecycle management. *The International Journal of Advanced Manufacturing Technology*, 81(1-4):667–684. 41

[Li et al., 2016] Li, W., Liu, K., Belitski, M., Ghobadian, A., and O'Regan, N. (2016). e-leadership through strategic alignment: An empirical study of small-and medium-sized enterprises in the digital age. *Journal of Information Technology*, 31(2):185–206. 18

[Liaw et al., 2002] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22. 54

[Lin, 2010] Lin, T. H. (2010). A comparison of multiple imputation with em algorithm and mcmc method for quality of life missing data. *Quality & quantity*, 44(2):277–287. 52

[Little, 1988] Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202. 51

[Louis, 2012] Louis, G. (2012). Pacte pour la compétitivité de l'industrie française. 13

[Luna, 2009] Luna, J. J. (2009). Metrics, models, and scenarios for evaluating phm effects on logistics support. In *Proceedings of Annual Conference of the Prognostics and Health Management Society*. 29

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA. 54

[Madanhire and Mbohwa, 2016] Madanhire, I. and Mbohwa, C. (2016). Enterprise resource planning (erp) in improving operational efficiency: Case study. *Procedia CIRP*, 40:225–229. 41, 104

[Madsen et al., 2016] Madsen, E. S., Bilberg, A., and Hansen, D. G. (2016). Industry 4.0 and digitalization call for vocational skills, applied industrial engineering, and less for pure academics. In *Proceedings of the 5th P&OM World Conference, Production and Operations Management, P&OM*. 40

[Malarvizhi and Thanamani, 2012] Malarvizhi, M. R. and Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development*, 5(1):5–7. 52

[Mallinson and Gammerman, 2003] Mallinson, H. and Gammerman, A. (2003). Imputation using support vector machines. *Department of Computer Science. Royal Holloway, University of London. Egham, UK*. 52

[Mariani et al., 2018] Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. (2018). Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*. 50, 51, 63

[Masood and Sonntag, 2020] Masood, T. and Sonntag, P. (2020). Industry 4.0: Adoption challenges and benefits for smes. *Computers in Industry*, 121:103261. ix, 18, 19, 21, 26

[Mathew et al., 2006] Mathew, A., Zhang, L., Zhang, S., and Ma, L. (2006). A review of the mimosa osa-eai database for condition monitoring systems. In *Engineering Asset Management*, pages 837–846. Springer. 25

[McGilvray, 2008] McGilvray, D. (2008). *Executing data quality projects: Ten steps to quality data and trusted information (TM)*. Elsevier. 45

[Mittal et al., 2018] Mittal, S., Khan, M. A., Romero, D., and Wuest, T. (2018). A critical review of smart manufacturing & industry 4.0 maturity models: Implications for small and medium-sized enterprises (smes). *Journal of manufacturing systems*, 49:194–214. 3, 13, 14, 15, 16, 18, 19, 20, 21, 26

[Moeuf, 2018] Moeuf, A. (2018). *Identification des risques, opportunités et facteurs critiques de succès de l'industrie 4.0 pour la performance industrielle des PME*. Theses, Université Paris Saclay (COmUE). 3, 16, 17, 19, 21, 104

[Moeuf et al., 2017] Moeuf, A., Lamouri, S., Pellerin, R., Eburdy, R., and Tamayo, S. (2017). Industry 4.0 and the sme: a technology-focused review of the empirical literature. In *7th International Conference on Industrial Engineering and Systems Management IESM*. 17, 19

[Moeuf et al., 2018] Moeuf, A., Pellerin, R., Lamouri, S., Tamayo-Giraldo, S., and Barbaray, R. (2018). The industrial management of smes in the era of industry 4.0. *International Journal of Production Research*, 56(3):1118–1136. 13, 16, 19

[Müller and Voigt, 2017] Müller, J. and Voigt, K. (2017). Industry 4.0—integration strategies for small and medium-sized enterprises. In *Proceedings of the 26th International Association for Management of Technology (IAMOT) Conference, Vienna, Austria*, pages 14–18. 20

[Myers, 2011] Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication methods and measures*, 5(4):297–310. 52

[Nakagawa and Freckleton, 2008] Nakagawa, S. and Freckleton, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in ecology & evolution*, 23(11):592–596. 52

[Nembrini et al., 2018] Nembrini, S., König, I. R., and Wright, M. N. (2018). The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718. 85

[Oh et al., 2018] Oh, H., Azarian, M. H., Cheng, S., and Pecht, M. G. (2018). Sensor systems for phm. *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*, pages 39–60. 42

[Omri, 2020] Omri, N. (2020). L'industrie 4.0 à la portée des pmes. *https://www.researchgate.net/publication/346651675*. 3, 15

[Omri et al., 2019] Omri, N., Al Masry, Z., Giampiccolo, S., Mairot, N., and Zerhouni, N. (2019). Data management requirements for phm implementation in smes. In *2019 Prognostics and System Health Management Conference (PHM-Paris)*, pages 232–238. IEEE. 3, 5, 13, 15, 20, 21, 22, 23, 27, 29, 39, 43, 47, 61

[Omri et al., 2020] Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni, N. (2020). Industrial data management strategy towards an sme-oriented phm. *Journal of Manufacturing Systems*, 56:23–36. ix, xiii, 3, 5, 16, 18, 21, 27, 28, 29, 30, 42, 43, 44, 45, 46, 47, 49, 54, 97

[Omri et al., 2021] Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni, N. (2021). Towards an adapted phm approach: Data quality requirements methodology for fault detection applications. *Computers in Industry*, 127:103414. x, 42, 43, 48, 54, 81, 82, 83, 104

[Orzes et al., 2018] Orzes, G., Rauch, E., Bednar, S., and Poklemba, R. (2018). Industry 4.0 implementation barriers in small and medium sized enterprises: A focus group study. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1348–1352. IEEE. 20

[Panahy et al., 2013] Panahy, P. H. S., Sidi, F., Affendey, L. S., Jabar, M. A., Ibrahim, H., and Mustapha, A. (2013). Discovering dependencies among data quality dimensions: A validation of instrument. *Journal of Applied Sciences*, 13(1):95–102. 46

[Pecht, 2009] Pecht, M. (2009). Prognostics and health management of electronics. *Encyclopedia of Structural Health Monitoring*. 22

[Pecht, 2010] Pecht, M. G. (2010). A prognostics and health management roadmap for information and electronics-rich systems. *IEICE ESS Fundamentals Review*, 3(4):4_25–4_32. 22

[Pérez et al., 2018] Pérez, J. D. C., Buitrón, R. E. C., and Melo, J. I. G. (2018). Methodology for the retrofitting of manufacturing resources for migration of sme towards industry 4.0. In *International Conference on Applied Informatics*, pages 337–351. Springer. 19

[Pipino et al., 2002a] Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002a). Data quality assessment. *Communications of the ACM*, 45(4):211–218. xiii, 46

[Pipino et al., 2002b] Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002b). Data quality assessment. *Commun. ACM*, 45(4):211–218. 45

[Rahman and Islam, 2011] Rahman, G. and Islam, Z. (2011). A decision tree-based missing value imputation technique for data pre-processing. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pages 41–50. 52

[Rauch et al., 2018] Rauch, E., Matt, D. T., Brown, C. A., Towner, W., Vickery, A., and Santiteerakul, S. (2018). Transfer of industry 4.0 to small and medium sized enterprises. *Transdisciplinary Engineering Methods for Social Innovation of Industry*, 4:63–71. 19

[Redman, 1997] Redman, T. C. (1997). *Data Quality for the Information Age*. Artech House, Inc., Norwood, MA, USA, 1st edition. 45

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. 66

[Ribeiro et al., 2018] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535. 66

[Rish et al., 2001] Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. 87

[Ross et al., 1990] Ross, B., Shultz, T., Silverstein, G., Wisniewski, E., et al. (1990). The influence of prior knowledge on concept acquisition: Experimental and computational results. 86

[Ross and Vitale, 2000] Ross, J. W. and Vitale, M. R. (2000). The erp revolution: surviving vs. thriving. *Information systems frontiers*, 2(2):233–241. 41, 104

[Roy and Dey, 2018] Roy, T. and Dey, S. (2018). Fault detectability conditions for linear deterministic heat equations. *IEEE control systems letters*, 3(1):204–209. 83

[Russom, 2008] Russom, P. (2008). Data governance strategies. In *Bus. Intell. J.*, volume 13, pages 13–14. 44

[Saaty, 1980] Saaty, T. (1980). Planning, priority setting, resource allocation. *The analytic hierarchy process*. 29

[Saenz de Ugarte et al., 2009] Saenz de Ugarte, B., Artiba, A., and Pellerin, R. (2009). Manufacturing execution system–a literature review. *Production planning and control*, 20(6):525–539. 42, 104

[Sahal et al., 2020] Sahal, R., Breslin, J. G., and Ali, M. I. (2020). Big data and stream processing platforms for industry 4.0 requirements mapping for a predictive maintenance use case. *Journal of Manufacturing Systems*, 54:138–151. 17

[Salles, 2006] Salles, M. (2006). Decision making in smes and information requirements for competitive intelligence. *Production Planning & Control*, 17(3):229–237. 15

[SBA2019, 2021] SBA2019 (2021). 2019 sba fact sheet (france). *http://ec.europa.eu/*, page (accessed: 22.01.2021). 13

[Schumacher et al., 2016] Schumacher, A., Erol, S., and Sihn, W. (2016). A maturity model for assessing industry 4.0 readiness and maturity of manufacturing enterprises. *Procedia Cirp*, 52:161–166. 17

[Schwab, 2017] Schwab, K. (2017). *The fourth industrial revolution*. Currency. 40

[SCODER, 2021] SCODER (2021). Scoder's website. *https://www.scoder.fr*, page (accessed: 10.02.2021). 10

[Sebastian-Coleman, 2012] Sebastian-Coleman, L. (2012). *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes. 39, 45

[Sevinc et al., 2018] Sevinc, A., Gür, Ş., and Eren, T. (2018). Analysis of the difficulties of smes in industry 4.0 applications by analytical hierarchy process and analytical network process. *Processes*, 6(12):264. 19

[Shawe-Taylor and Cristianini, 2002] Shawe-Taylor, J. and Cristianini, N. (2002). On the generalization of soft margin algorithms. *IEEE Transactions on Information Theory*, 48(10):2721–2735. 73

[Shi et al., 2020] Shi, D., Lee, T., Fairchild, A. J., and Maydeu-Olivares, A. (2020). Fitting ordinal factor analysis models with missing data: A comparison between pairwise deletion and multiple imputation. *Educational and psychological measurement*, 80(1):41–66. 52

[Sidi et al., 2012] Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., and Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management*, pages 300–304. IEEE. 45

[Singh et al., 2007] Singh, R. K., Garg, S. K., Deshmukh, S., and Kumar, M. (2007). Modelling of critical success factors for implementation of amts. *Journal of Modelling in Management*. 15

[Singh et al., 2016] Singh, S., Ribeiro, M. T., and Guestrin, C. (2016). Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*. 66

[Spinner et al., 2019] Spinner, T., Schlegel, U., Schäfer, H., and El-Assady, M. (2019). explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26(1):1064–1074. 61

[Srinivasan et al., 2020] Srinivasan, S., Cawi, E., Hyman, J., Osthus, D., Hagberg, A., Viswanathan, H., and Srinivasan, G. (2020). Physics-informed machine learning for backbone identification in discrete fracture networks. *Computational Geosciences*, 24:1429–1444. 61

[Subrahmanya, 2015] Subrahmanya, M. B. (2015). Innovation and growth of engineering smes in bangalore: why do only some innovate and only some grow faster? *Journal of Engineering and Technology Management*, 36:24–40. 14, 21

[Sun et al., 2007] Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378. 50

[Suri et al., 2019] Suri, N. R., Athithan, G., et al. (2019). Research issues in outlier detection. In *Outlier Detection: Techniques and Applications*, pages 29–51. Springer. 52

[Takata et al., 2004] Takata, S., Kirnura, F., van Houten, F. J., Westkamper, E., Shpitalni, M., Ceglarek, D., and Lee, J. (2004). Maintenance: changing role in life cycle management. *CIRP annals*, 53(2):643–655. 23

[Tan et al., 2020] Tan, S., Soloviev, M., Hooker, G., and Wells, M. T. (2020). Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pages 23–34. 65

[Tang et al., 2015] Tang, X.-m., Yuan, R.-x., and Chen, J. (2015). Outlier detection in energy disaggregation using subspace learning and gaussian mixture model. *Int. J. Control Autom.*, 8(8):161–170. 54

[Tao et al., 2018] Tao, F., Qi, Q., Liu, A., and Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, 48:157–169. 40, 41

[Terziovski, 2010] Terziovski, M. (2010). Innovation practice and its performance implications in small and medium enterprises (smes) in the manufacturing sector: a resource-based view. *Strategic Management Journal*, 31(8):892–902. 20

[Thurston and Lebold, 2001] Thurston, M. and Lebold, M. (2001). Standards developments for condition-based maintenance systems. Technical report, PENNSYLVANIA STATE UNIV UNIVERSITY PARK APPLIED RESEARCH LAB. 25

[Toth et al., 2018] Toth, N., Ladányi, R., and Garamvölgyi, E. (2018). Elaborating industry 4.0 compatible dss for enhancing production system effectiveness. In *IOP Conference Series: Materials Science and Engineering*, volume 448, page 012040. IOP Publishing. 17

[Trappey et al., 2017] Trappey, A. J., Trappey, C. V., Govindarajan, U. H., Chuang, A. C., and Sun, J. J. (2017). A review of essential standards and patent landscapes for the internet of things: A key enabler for industry 4.0. *Advanced Engineering Informatics*, 33:208–229. 17

[Tsai et al., 2009] Tsai, C.-h., Chang, L.-c., and Chiang, H.-c. (2009). Forecasting of ozone episode days by cost-sensitive neural network methods. *Science of the Total Environment*, 407(6):2124–2135. 50, 63

[Tso and Yau, 2007] Tso, G. K. and Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768. 86

[Van Buuren and Oudshoorn, 1999] Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO. 52

[Van de Vrande et al., 2009] Van de Vrande, V., De Jong, J. P., Vanhaverbeke, W., and De Rochemont, M. (2009). Open innovation in smes: Trends, motives and management challenges. *Technovation*, 29(6-7):423–437. 15

[Vogl et al., 2014] Vogl, G. W., Weiss, B. A., and Donmez, M. A. (2014). Standards for prognostics and health management (phm) techniques within manufacturing operations. Technical report, National Institute of Standards and Technology Gaithersburg United States. 23, 26, 39

[Vogl et al., 2019] Vogl, G. W., Weiss, B. A., and Helu, M. (2019). A review of diagnostic and prognostic capabilities and best practices for manufacturing. *Journal of Intelligent Manufacturing*, 30(1):79–95. 23

[von Rueden et al., 2019] von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., et al. (2019). Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems. *arXiv preprint arXiv:1903.12394*. 59, 60, 61, 62

[Wang et al., 2007] Wang, C., Walker, E., and Redmond, J. (2007). Explaining the lack of strategic planning in smes: The importance of owner motivation. 15

[Wang et al., 2018] Wang, J., Ma, Y., Zhang, L., Gao, R. X., and Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 48:144–156. 12

[Wang, 2013] Wang, K.-S. (2013). Towards zero-defect manufacturing (zdm)—a data mining approach. *Advances in Manufacturing*, 1(1):62–74. 21

[Wang and Strong, 1996] Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33. 45

[Wang et al., 2016] Wang, Y., Wang, G., and Anderl, R. (2016). Generic procedure model to introduce industrie 4.0 in small and medium-sized enterprises. In *Proceedings of the world congress on engineering and computer science*, volume 2. 19

[Wang and Witten, 2002] Wang, Y. and Witten, I. H. (2002). Modeling for optimal probability prediction. 90

[Wank et al., 2016] Wank, A., Adolph, S., Anokhin, O., Arndt, A., Anderl, R., and Metternich, J. (2016). Using a learning factory approach to transfer industrie 4.0 approaches to small-and medium-sized enterprises. *Procedia Cirp*, 54:89–94. 19

[Waurzyniak, 2015] Waurzyniak, P. (2015). Why manufacturing needs real-time data collection. *Manufacturing Engineering*, pages 53–61. 14

[Wolberg and Mangasarian, 1990] Wolberg, W. H. and Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the national academy of sciences*, 87(23):9193–9196. 86

[Würtz and Kölmel, 2012] Würtz, G. and Kölmel, B. (2012). Integrated engineering–a sme-suitable model for business and information systems engineering (bise) towards the smart factory. In *Working Conference on Virtual Enterprises*, pages 494–502. Springer. 19

[Xu et al., 2015] Xu, J., Wang, Y., and Xu, L. (2015). Phm-oriented sensor optimization selection based on multiobjective model for aircraft engines. *IEEE Sensors Journal*, 15(9):4836–4844. 29

[Xu et al., 2018] Xu, M., Jin, X., Kamarthi, S., and Noor-E-Alam, M. (2018). A failure-dependency modeling and state discretization approach for condition-based maintenance optimization of multi-component systems. *Journal of manufacturing systems*, 47:141–152. 23

[Xu, 2012] Xu, X. (2012). From cloud computing to cloud manufacturing. *Robotics and computer-integrated manufacturing*, 28(1):75–86. 17

[Yang et al., 2009] Yang, D., Rundensteiner, E. A., and Ward, M. O. (2009). Neighbor-based pattern detection for windows over streaming data. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 529–540. 54

[Yeh and Lien, 2009] Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480. 86

[Yeh et al., 2009] Yeh, I.-C., Yang, K.-J., and Ting, T.-M. (2009). Knowledge discovery on rfm model using bernoulli sequence. *Expert Systems with Applications*, 36(3):5866–5871. 90

[Yew Wong and Aspinwall, 2004] Yew Wong, K. and Aspinwall, E. (2004). Characterizing knowledge management in the small business environment. *Journal of Knowledge management*, 8(3):44–61. 21

[Zaveri et al., 2016] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93. 44

[Zemouri et al., 2018] Zemouri, R., Omri, N., Devalland, C., Arnould, L., Morello, B., Zerhouni, N., and Fnaiech, F. (2018). Breast cancer diagnosis based on joint variable selection and constructive deep neural network. In *2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME)*, pages 159–164. IEEE. 63, 67

[Zemouri et al., 2019] Zemouri, R., Omri, N., Fnaiech, F., Zerhouni, N., and Fnaiech, N. (2019). A new growing pruning deep learning neural network algorithm (gp-dlnn). *Neural Computing and Applications*, pages 1–17. 63, 67, 72, 86

[Zheng et al., 2016] Zheng, Z., Cai, Y., and Li, Y. (2016). Oversampling method for imbalanced classification. *Computing and Informatics*, 34(5):1017–1037. 50, 51

[Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929. 66

Abstract:

Data technologies' emergence brings many developments in the industrial domain and pushed companies on the road to digitization. Nevertheless, Industry 4.0, Big data, and artificial intelligence are concepts more associated with large companies. One of the current challenges for SMEs is a reflection on these concepts' appropriation to avoid a competitive gap with large groups. In this context, the SCODER company is leading a voluntarist digitization strategy. One of this strategy's axes corresponds to scientific collaboration with the Femto-st institute and the ENSMM of Besançon (CIFRE thesis). The objective was to develop a methodology for Industry 4.0 integration within SMEs. Thus, a bibliographic study was conducted to identify the barriers that limit the Industry 4.0 implementation within SMEs. Based on these barriers, an adapted prognostics and health management (PHM) approach is proposed as a solution to implement Industry 4.0 within SMEs. This approach is applied in the SCODER company, which permits identifying the data quality management issue as the most critical research gaps. Thus it is considered as the backbone of this thesis work. However, traditional data quality improvement techniques are not applicable in the case of SMEs. Thus, a knowledge-based data quality improvement framework is proposed to deal with this problem. Firstly, a generic formalization of the operators' know-how is introduced to improve data quality and improve the data analysis results. In a second step and to ensure efficient knowledge management, the obtained analysis results are explained and used to enrich the human knowledge applied to improve data quality in future analysis tasks. Data management can generate a non negligible cost for SMEs. For that, an empirical metric is developed to quantify the data quality impact on the PHM results and thus identify requirements on data quality to satisfy each objective. Based on this metric, different strategies are proposed to improve performance at the right cost. To facilitate its implementation, all the steps of the proposed methodology have been encapsulated in SCODER Data System (DS2) software.

Keywords: Prognostics and health management, Data analysis, Machine learning , SMEs, Data quality, Explainable AI.

Résumé :

L'émergence des technologies de données apporte de nombreux développements dans le domaine industriel et a poussé les entreprises sur la voie de la numérisation. Néanmoins, Industrie 4.0, Big data et intelligence artificielle sont des concepts plutôt associés aux grandes entreprises. L'un des enjeux actuels pour les PME est une réflexion sur l'appropriation de ces concepts pour éviter un écart de compétitivité avec les grands groupes. Dans ce contexte, l'entreprise SCODER mène une stratégie de digitalisation volontariste. Un des axes de cette stratégie correspond à une collaboration scientifique avec l'institut Femto-st et l'ENSMM de Besançon (thèse CIFRE). L'objectif était de développer une méthodologie d'intégration de l'industrie 4.0 au sein des PME. Ainsi, une étude bibliographique a été menée pour identifier les barrières qui limitent l'implémentation de l'industrie 4.0 dans les PME. Sur la base de ces obstacles, une approche adaptée de la gestion du pronostic et de la santé (PHM) est proposée comme solution pour mettre en œuvre l'industrie 4.0 dans les PME. Cette approche est appliquée dans l'entreprise SCODER, ce qui permet d'identifier la question de la gestion de la qualité des données comme la plus critique des lacunes de recherche. Elle est donc considérée comme l'épine dorsale de ce travail de thèse. Cependant, les techniques traditionnelles d'amélioration de la qualité des données ne sont pas applicables dans le cas des PME. Ainsi, un cadre d'amélioration de la qualité des données basé sur la connaissance est proposé pour traiter ce problème. Dans un premier temps, une formalisation générique du savoir-faire des opérateurs est introduite pour améliorer la qualité des données et les résultats de l'analyse des données. Dans un deuxième temps et pour assurer une gestion efficace des connaissances, les résultats d'analyse obtenus sont expliqués et utilisés pour enrichir les connaissances humaines appliquées pour améliorer la qualité des données dans les futures tâches d'analyse. La gestion des données peut générer un coût non négligeable pour les PME. Pour cela, une métrique empirique est développée pour quantifier l'impact de la qualité des données sur les résultats du PHM et ainsi identifier les exigences de qualité des données pour satisfaire chaque objectif. Sur la base de cette métrique, différentes stratégies sont proposées pour améliorer les performances au bon coût. Pour faciliter sa mise en œuvre, toutes les étapes de la méthodologie proposée ont été encapsulées dans le logiciel SCODER Data System (DS2).

Mots-clés : Prognostics and health management, Analyse de données, Machine learning, PME, Qualité de données, IA explicable.

SPIM

■ École doctorale SPIM 16 route de Gray F - 25030 Besançon cedex
■ tél. +33 (0)3 81 66 66 02 ■ ed-spim@univ-fcomte.fr ■ www.ed-spim.univ-fcomte.fr

UFC UNIVERSITÉ DE FRANCHE-COMTÉ