

Politecnico di Torino



COLLEGIO DI INGEGNERIA ELETTRONICA, DELLE TELECOMUNICAZIONI E FISICA

MSC IN ICT FOR SMART SOCIETIES ENGINEERING

Operation variability and line balancing management in a high quality assembly line by machine learning analytics

Supervisor

Prof. Andrea ACQUAVIVA

Author

Gaetano MOCERI

Supervisor at FCA World Class Manufacturing Development Center

Ing. Francesco CANUTO

April 2018

Alla mia famiglia.

Abstract

Variability of assembly operations (MURA) in a high quality production line is one of the main sources of losses in assembly operations. Typically it is caused by difficult operations which the operator cannot always perform in the same time. Common causes are product tolerances, method weakness, weak tooling design or balancing not efficient. Moreover, on top of variability in timing of the operations in a workplace caused by difficult operations one must add that caused by vehicle and logistic complexity (different combinations of product features and optional). A high level of variability negatively affects the stability and robustness of line balancing since operators could be unable to perform the required operations in the expected time and therefore potentially exceed the set tack-time. The result of such situation is the generation of multiple micro stoppages (loss in line productivity), quality issues due to an increase in stress of the operator and workplaces non accessible to everyone. However, identifying such losses is difficult since it requires multiple (usually at least thirty records a operator) and accurate measurements of the duration of operations. Traditionally, Industrial Engineering identifies those stations where the variability is suspected to generate problems and spends a considerable effort in measuring the real duration of operations during a shift or throughout different shifts. Even if this approach is accurate and effective, it is very time-consuming especially when it comes to identify variability and can therefore not be used to cover all stations. Conversely, in the new generation of Manufacturing Execution System (MES) of FCA (Fiat Chrysler Automobiles) Plants, timestamps of various operations are stored. Therefore, the aim of this work is trying to give a different approach to the Mura analysis, different from the traditional one, thanks to the new Industry 4.0 paradigms, through building a tool able to perform statistical analysis able to describe operations' features and detect the variability contributors, by using also machine learning techniques.

Acknowledgements

The experience done into FCA was a great chance to me to touch the business environment and to work into a huge and very important company. I had the opportunity of working with a lot of great and important professional figures that gave me a lot of tips for my future career in the labour market. One of these is undoubtedly Ing. Francesco Canuto, which is a very skilled person and a great leader within the team. I would like to thank him for his precious advice and for the opportunity that WCM office gave me. I want to thank also Andrea Bellagarda, for his guidance for what concerns training sessions, presentations and meeting with stakeholders involved into the project. I want to thank the WPI team, in particular Alessandro Lacalaprice and Vincent Ruelle, because they let me participate to very important meetings within Industry 4.0 projects in FCA and they gave me the opportunity to know interesting opportunities for the future of my career.

Of course, I want also to thank prof. Acquaviva and prof. Macii, which allowed me to start this internship giving me the chance of working in one of the most important FCA offices.

Contents

List of Tables	VI
List of Figures	VII
1 Introduction	1
1.1 Introduction	1
1.2 Thesis goal	2
2 Fiat Chrysler Automobiles and the Word Class Manufacturing	4
2.1 Background	4
2.2 Industry 4.0 and digital revolution	5
2.2.1 Digital in FCA manufacturing	9
2.3 FCA plant structure	12
2.4 World Class Manufacturing	14
2.4.1 Introduction to WCM	14
2.4.2 WCM structure	17
2.4.3 Workplace Organization sub-pillar	18
Basic principles	19
Step 1	21
Step 2	22
2.4.4 Mura	23
3 Analysis	27
3.1 Mura analysis: premise	27
3.2 Mura analysis: state of the art	29
3.2.1 Weakness of the standard Mura analysis and advantages of Machine Learning	31

3.3	Automatic Mura analysis	31
3.3.1	Statistical Analysis	33
	Data cleaning	35
	Normality tests	39
	Outlier detection	43
	Ordered operation	47
	Autocorrelation analysis	50
	Data conversion	52
	Dataset consistency	52
3.3.2	Contributors detection	53
	Classification labeling	54
	Machine learning algorithms	57
3.3.3	Upsampling of the whole dataset and using the same train and test dataset	61
3.3.4	Upsampling of the whole dataset and split train and test datasets	62
3.3.5	Upsampling only train dataset	64
	Separating the decision regions	64
4	Summary	67
5	Future works	70
	Bibliography	72
	List of Acronyms	77

List of Tables

3.1	The table below summarizes the classification performances for the upsampled dataset.	63
3.2	The table below summarizes the classification performances for the upsampled training set.	64

List of Figures

2.1	Smart factory model. The actual strength is the point number 3, where data gathered are managed, computed using state-of-the-art methodologies and transformed into useful information to the company business.	5
2.2	Industrial revolutions main changes [39]	6
2.3	General concept of a smart manufacturing enterprise [21].	7
2.4	Caption	9
2.5	Digital manufacturing for FCA company. On the left the enabling technologies exploited in plants. On the right the main fields of application of such technologies.	10
2.6	The graph shows a conceptual model of the global structure of a smart factory. . .	11
2.7	Assembling of the body metal components	12
2.8	The assembly line of Melfi plant	13
2.9	Global organization of an FCA plant.	14
2.10	Standard Kaizen: used for the implementation of a project aiming to the continuous improvement.	15
2.11	WCM temple structure: columns represent technical pillars. Basis represent managerial pillars.	18
2.12	Figure shows 10 operators doing 10 operations each. The arrow highlights the bottleneck operations and "forces" the time cycle to be very close to the takt time.	20
2.13	Actual C-Matrix from cost deployment. It shows as the main losses sources are NVAA, logistics and unbalancing: typical human related operations.	21
2.14	Total losses per WCM pillar. The first one is Workplace Organization, highlighting its importance in the plant business.	21
2.15	5S application picture	22
2.16	The figure shows the composition of a Time Cycle for an operator.	23

2.17	Effect on the nominal time of the results of a Mura analysis	25
2.18	Re-organization of the labor intensive process due to application of Step 2 of WO pillar	25
2.19	New line balancing coming from the elimination of NVA actions. Operators have been displaced to other tasks, with economic benefits for the plant.	26
2.20	Actual Key Performance Indicator of the implementation of step 2 in a Fiat Power- Train plant.	26
3.1	Operations tag for the specified workplace, with each single operation's description	28
3.2	Job Element Sheet (JES): shows in a very detailed fashion the exact sequence of the operation to be performed. The example shows, also with a graphical aid for the sake of clarity, how to fix the window to the body of the door.	29
3.3	A result of the 30 cycle recording on a histogram for the "Operator A"	30
3.4	Distribution of operation times. In x axis, the mean time. In y axis the percentage of samples.	30
3.5	Example of process of a single workplace. Manuel operations are not considered in MES detection. In this case the line stops due to a delay coinciding with the last quality report operation (in red).	32
3.6	Sample of MES extraction from AGAP plant.	34
3.7	Box plot example for a Gaussian distribution.	36
3.8	Visualization plot operation 2: scatter plot, box plot and times histogram	39
3.9	Visualization plot operation 6: scatter plot, box plot and times histogram	40
3.10	Visualization plot operation 5: scatter plot, box plot and times histogram	40
3.11	Summary of normality test for the available operations.	42
3.12	Conceptual schema for data visualization [24]	43
3.13	Scatter plots for two operations: outliers are labeled by red dots.	45
3.14	Box plot per each execution day for operation 3.	45
3.15	Box plot of operation clustered per execution day.	46
3.16	Box plot of operation clustered per hour of the day. It shows a huge variability during the early phase of the shift.	46
3.17	Scatter plots for operation number 5 for the two teams involved.	48
3.18	Caption	48
3.19	MES relevant operations for a single workplace. X axis the operations, Y axis the execution time.	50

3.20	Example of dataframe with the updated features	52
3.21	PCA applied on the dataframe of operation number 4	53
3.22	Qualitative example of classification for a (supposed) Gaussian distribution of operation times.	54
3.23	Diagrammatic representation by the producers of LightGBM to explain the difference between level and leaf wise algorithms.	59
3.24	Operation 2: LightGBM performance analysis for training and testing on the same dataset.	62
3.25	Operation 2: LightGBM performance analysis for upsampled dataframe. Training and testing phase made by splitting dataset.	63
3.26	Scatter plot that shows the separation of decision regions. Time is logarithmic in order to allow a more balanced division.	65
3.27	Caption	65

Listings

3.1	Data opening and creation of the data structure	34
3.2	Code showing the very first step of datasets cleaning	35
3.3	Method that plots scatter plot box plot and times histograms for every team presents in the dataframe	37
3.4	Method used to add an "outlier" column and flag the entries as outlier or not. . . .	44
3.5	Key method to evaluate whenever the two teams use to switch the operation execu- tion with another one	49
3.6	Methods that perform classification based on the concept of "reference time"	55
3.7	This method computes the performance of the classification	56

Chapter 1

Introduction

1.1 Introduction

This master thesis is part of the new trend of car makers and manufacturers in general of moving technologies, methodologies and tools towards digital and Industry 4.0 concepts. In particular, the focus of these new paradigms applied to cars production involves the utilization of very new academical knowledge to existing plants and Information Technologies infrastructure in order to find useful insights to improve performances, enlarge the business and reduce wastes and losses.

At the moment, FCA plants have reached a very high degree of digitalization in which there are a lot of softwares that hold different functionalities, e.g. robot management, production flow control, logistics management, and collect data from different data sources. Unfortunately, very often, data collected are unused and platforms are independent each other causing a serious waste of opportunity in terms of analysis. Hence, the guideline for the next future of the company is trying to build a platform in which the whole production pours every kind of data with the purpose of aggregating them and then performing more in-depth analysis involving Big Data, Machine Learning and Artificial Intelligence techniques. The final goal is reaching a condition of Data Driven Decision Making (DDDM) which is an approach to business company governance that values decisions that can be backed up with verifiable data. The success of the data-driven approach is reliant upon the quality of the data gathered and the effectiveness of its analysis and interpretation [6].

Within this scenario, the very early stage of the implementation of the digital revolution deals with gathering some practical use cases with the purpose of testing some innovative solution to real industrial problems and investigate the potential advantages of using the newest solutions

and technologies, e.g. virtual reality, augmented reality, Internet of Things devices, etc. This innovation path should be the test bench where digital methodologies mix with manufacturing process underlying the benefits and improvements with respect to the traditional one. These use cases, are typically proposed by people working in plants and white collars expose a practical problem that could be solvable thanks to digital technologies that do not exist yet in FCA's systems.

1.2 Thesis goal

The goal of this Master Thesis is trying to give a methodological solution to an use case proposed by "Avv. Giovanni Agnelli" Plant (AGAP) about the analysis of the variability (Mura) of labor intensive operations¹ that are severely affected by NVA Activities², representing a bunch of problems for either plant and company business, workers safety and product quality. The scenario of the analysis is a high quality assembly line. High quality reflects the fact that the time cycle is very high and the number of operations made by workers is significantly higher than a traditional assembly line. In particular, at the moment, the variability of operations is very difficult to be taken into account because it deals with off-line and very expensive analysis where a worker should film the supposed most critical station. timing the operation times. If the number of critical operations is very high, the analysis might be very expensive in terms of time and money.

This work aims to provide a tool able to automate the Mura analysis and detection, investigate the statistical distribution of operation times for different operations and operators; try to find correlations between features using statistical tools and machine learning algorithms and find the main contributors, compute some statistical parameters in order to extract new KPIs (Key Performance Indicators) useful to keep under control labor intensive areas and detecting the most critical stations or operations and the possible factors that cause variability and solving them to improve productivity performances. Two different machine learning algorithms have been tested with the available dataset: performances have been evaluated and the actual potential was brought to the attention of the plants stakeholders.

¹Labor intensive refers to a process or industry that requires a large amount of human labor to produce its goods or services. On the other side, capital intensive, refers typically to automated and robotic processes.

²Non-Value Added Activities add costs to a product but do not add any value to the realization or management of itself.

Operations data are stored in MES (Manufacturing Execution System), which is a computerized system that stores features of the labor intensive operations. In this work, the data source is represented by nine MES extractions into Excel files containing just two weeks of operations made in "Avv. Giovanni Agnelli" Plant (AGAP). Files contain just temporal data and information related to operators and teams.

The core of this work is basically extracting useful information from existing data. Hence, a data analysis activity has been conducted, performed with Python programming language because this language is going to establish itself as one of the most popular languages for scientific computing, thanks to its high-level interactive nature and its huge ecosystem of scientific libraries, for the statistical analysis, machine learning algorithms and results visualization. It is an appealing choice for algorithmic development and data analysis [11] in trend with the very last state-of-the-art research fields. The code is developed within the Anaconda platform, using both Spyder IDE and Jupyter Notebook.

Chapter 2

Fiat Chrysler Automobiles and the Word Class Manufacturing

2.1 Background

This thesis is the outcome of an internship experience in the World Class Manufacturing (WCM) Development Center of Fiat Chrysler Automobiles Group.

Since this work is inserted in a sort of digital revolution in the manufacturing field, is important to explain the main concepts behind the methodologies and tools used to build an efficient and productive car making process applied by the company. WCM is the cornerstone of FCA manufacturing and with the advent of Industry 4.0, it is going to implement tools more advanced and innovative to solve traditional and difficult problems and improving manufacturing methodology. Hence, a solid introduction to the Industry 4.0 context should be discussed. It is characterized by innovative Information and Communication Technologies (ICT) applications with the final consequence to reach the status of "*smart factory*" (shown as a conceptual model in the figure 2.3) able to change every aspect of manufacturing management.

This work is going to explore some practical case studies that should introduce the combination of the traditional manufacturing methodologies with advanced Industry 4.0 applications, by not distorting the existing structure of the plant, but simply *exploiting systems and data already present in IT systems and try to use them following innovative knowledge and tools as Machine Learning and Artificial Intelligence*

The WCM Development Center [18] is the bridge between innovation, research and development

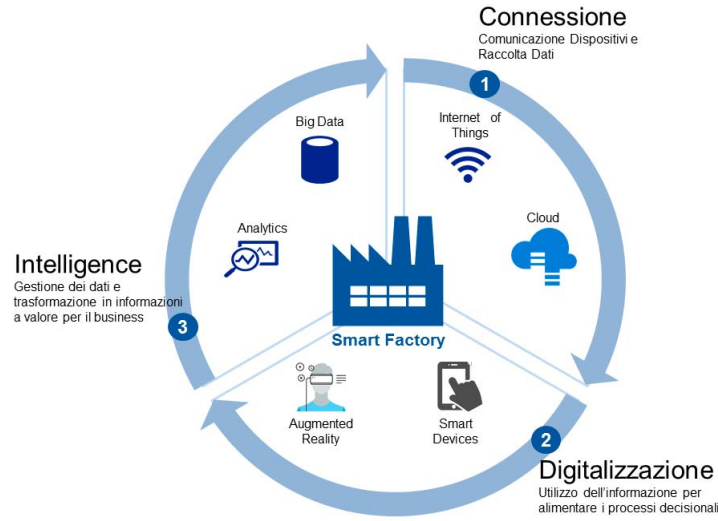


Figure 2.1: Smart factory model. The actual strength is the point number 3, where data gathered are managed, computed using state-of-the-art methodologies and transformed into useful information to the company business.

of the methodologies and application at Group plants. The Center develops methods and tools to support the evolution of WCM. The Center's activities include management of a large number of projects, training and coaching, support with implementation and development of best practices. It is also responsible for people development through specialized task forces, planning workshops, training events and web-based events.

The next section tells about the already mentioned Industry 4.0 paradigms and concepts and gives an idea of how this revolution is going to transform and change industrial processes and manufacturing.

2.2 Industry 4.0 and digital revolution

Throughout history, industrial revolutions characterized precise moments in which human life widely change from technological, socioeconomic and cultural point of view. The first industrial revolution was the major step in human history; in particular, regarding manufacturing technology, where production methodologies moved from human powered technology to machines. This lead to both factories emerging as well as new ways of processing old and new materials and improved water power and the use of the steam-engine. The second industrial revolution introduced the power of electricity, chemical products, oil and the combustion engine and lead to a shift towards

a new economy. If during the first industrial revolution consequences were slow, with the second one they were very faster and society radically changed, from the humblest to the richest, with the birth of new jobs, as laborer and industrial capitalist.

The third industrial revolution refers to a modern digital revolution. It started after the Second World War and during Cold War, where scientific discoveries and new technologies, especially in telecommunications field, represented the actual strength of Nations. For example, in 1969, the American Department of Defense built the first working telematics network able to allow communication among remote places for security and research reasons, called ARPANET¹, that was going to be the ancestor of the Internet of nowadays. The effect of the third industrial revolution on industry materialized with the advent of computers that boosted the beginning of automation, when robots and machines began to replace human workers in manufacturing processes, assembly lines and general management like transports, logistics, etc.

Nowadays, the very last trend in industry is the so called "*Industry 4.0*". Industry 4.0 was proposed by the German Government, in which the essence is mixing the Internet with the manufacturing. Industry 4.0 describes the future of manufacturing and will be established over the Internet and Information and Communication Technologies (ICT) based on innovative interactive platforms, the Internet of Things (IoT), industrial Internet of Things (IIoT), clouds, Artificial Intelligence (AI) and Machine Learning (ML) [7] (Figure 2.2). The fourth industrial revolution takes the

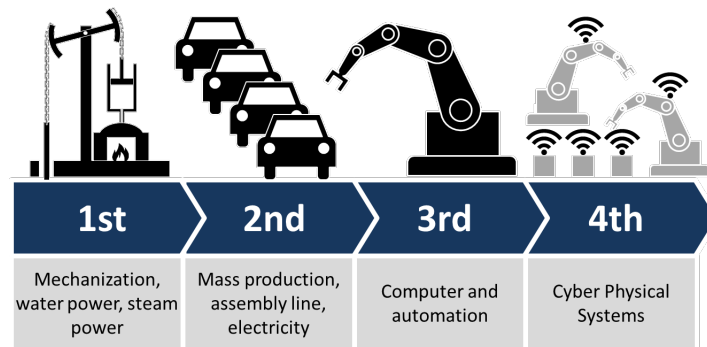


Figure 2.2: Industrial revolutions main changes [39]

automation of manufacturing processes to a new level by introducing those customized and flexible mass production technologies. The main consequence is that machines will operate independently, or cooperate with humans in creating a customer-oriented production field that constantly works

¹ ARPANET was an early packet switching network and the first network to implement the protocol suite TCP/IP. Both technologies became the technical foundation of the Internet

on maintaining itself, giving the birth to the so called "*smart factory*" (figure 2.3). The machine rather becomes an independent entity that is able to collect data, analyze it, monitoring physical processes and taking decisions upon it. This becomes possible by introducing different new concepts as self-optimization, self-cognition, and self-customization into the industry. The manufacturers would be able to communicate with computers rather than operate them.

Andrew Kusiak gives general concept of a smart manufacturing enterprise in his publication [21]. Basically, the main transformation in manufacturing is the birth of the Cyber-Physical System (CPS) able to provide a lot of new functionality and management methods.

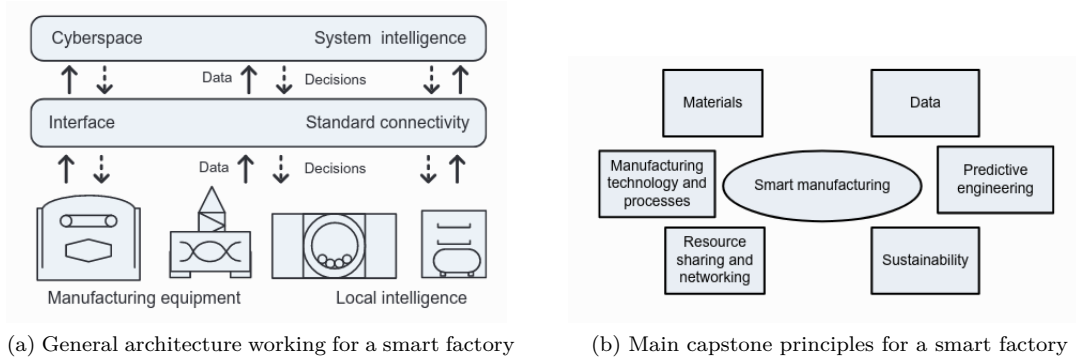


Figure 2.3: General concept of a smart manufacturing enterprise [21].

Every existing mechatronics system has been transformed into a CBS, allowing humans to perform complex tasks that require a minimum of suitability and specialized knowledge. Basically, these systems act as industrial production integrators of computing and management of the physical process happening in a plant, interacting in a dynamics continuous time and discrete events.

Cyber-Physical System, hence, connects the virtual space with the physical reality, which integrates computing, communication and storage capabilities, and can be real-time, reliable, secure, stable and efficient operation. The core concept of the Cyber-Physical System is: computation, communication and control, to achieve collaborative and real-time interaction between the real (physical) world and the information world through feedback loops of the interaction between computational processes and physical processes [29] [35].

Basically, there are four general design patterns within Industry 4.0. These are used as guidelines by companies to implement industry 4.0 scenarios and are summarized as follows:

- **Interoperability** Represents the capability of machines, electronic devices, sensors and software to connect and communicate each other via Internet or Internet of Things.

- **Information transparency** Deals with the free access to data, which would allow to create a virtual copy of the physical world by enhancing digital plant models with smart sensors data with huge advantages in simulations. This requires the aggregation of different data coming from different sources to higher-value context information in order to build and validate the model.
- **Technical assistance** Consists of supporting human activities by aggregating and visualizing information comprehensibly. Then, it deals with the ability of CPS to physically support humans by conducting a range of tasks that are difficult, unpleasant, too exhausting or unsafe for humans.
- **Decentralized decisions** The ability of CPS to make decisions on their own starting from autonomously performed analysis and to perform their tasks as independently as possible. This leads to the ideal situation of data-driven decision making, where data provides the necessary insights to properly move towards business decisions.

But as with any major shift and changing, there are a lot of challenges inherent in adopting this model like: data security issues that are hugely increased by integrating new systems and more access to those systems, a high degree of reliability and stability needed for successful of cyber-physical communication that needs a robust and secure physical infrastructure, maintaining the integrity of the production process with less human interventions with the consequent saving on workers.

By the way, there is a systemic lack of experience and manpower to create and implement these systems (especially in Italian companies), without mentioning a general reluctance from stakeholders, investors and high company management, even if there are a bunch of advantages in using Industry 4.0 practices they are not totally aware of them². Globally, when it is about increasing the degree of digitalisation of the company, managers agree the changing. But when the proposal is renewing the business model, which is actually the central theme of Industry 4.0, management is reluctant to move in that direction [25].

However, for sake of clarity, smart manufacturing is not about the degree of automation of the manufacturing floor; it is about autonomy, evolution, simulation and optimization of the manufacturing enterprise. The scope and time horizon of the simulation and optimization will depend on

²A study from Politecnico di Milano shows how Italy is still very backlog in Industry 4.0 theme: the 23% of the managers interviewed declared that they have no idea about these themes.

the availability of data and tools. The level of "*smartness*" of a manufacturing enterprise will be determined by the degree to which the physical enterprise has been reflected and invested in the cyber space.

Before going in the detail of Industry 4.0 and the description of the scope of this work and the reached results, is needed to briefly show how the company is moving in the Industry 4.0 scenario and what is the direction it is walking towards.

2.2.1 Digital in FCA manufacturing

Within this scenario of innovation and new ideas, how is FCA going to act in manufacturing innovation? Actually, the digital manufacturing field was born a lot of years ago, when the personal computers and electronic devices started to be controllable from remote: that was the birth of large scale automation. The figure 2.4 shows the WCM evolution steps, from its birth to the implementation of "*digital WCM*", which means that the new digital methodologies have heavily embraced WCM [13].

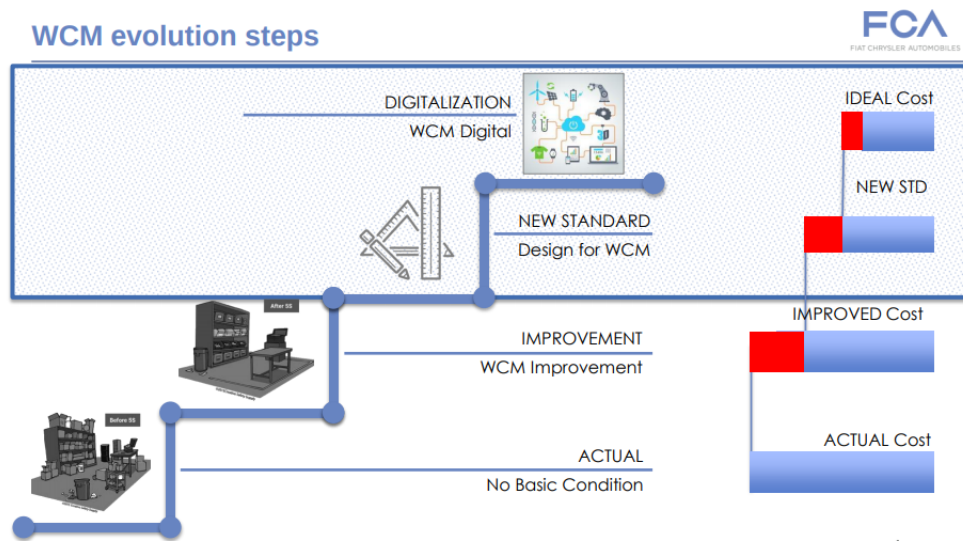


Figure 2.4: Caption

Today, with Industry 4.0 there are not only automated and very smart processes, but also, thanks to a lot of new concepts like machine learning, artificial intelligence, business intelligence, companies like FCA are able to find insights from data in order to improve their own business in advance. Machine learning's core technologies align well with the complex problems that manufacturers face every day: from striving to keep supply chains operating efficiently to produce customized, built-to-order products on time, machine learning algorithms have the potential to

bring greater predictive and analysis accuracy to every phase of the production.



Figure 2.5: Digital manufacturing for FCA company. On the left the enabling technologies exploited in plants. On the right the main fields of application of such technologies.

The figure 2.5 synthetically explains how the company is behaving with respect to the Industry 4.0: on the left side there are the main enabling technologies exploited for the continuous improvement projects, while on the right side, there is a list of main project areas, where the digital projects are going to be implemented. In this work, the focus is placed into big data and machine learning analytics: they are fundamental to the digital revolution and in realizing systems with intelligent behavior [22].

Many of the algorithms used within this scenario being developed are iterative, designed to learn continually and seek optimized outcomes. In order to explain and clarify the opportunity that algorithms may bring to manufacturers, is reported the following list [8]:

- Increasing production capacity while lowering material consumption: smart manufacturing systems designed to capitalize on predictive data analytics and machine learning have the potential to improve yield rates at the machine, production cell, and plant levels.
- Providing more relevant data such that operations, finance and supply chain teams can better manage factories and demand-side constraints.
- Improving preventive maintenance and maintenance in general terms. It would prevent useless repairing procedures and line stoppages, with the consequent reduction of productivity.

Integrating machine learning databases, apps, and algorithms into cloud platforms are becoming pervasive. The following graphic (figure 2.6) illustrates how machine learning is integrated from a conceptual point of view.

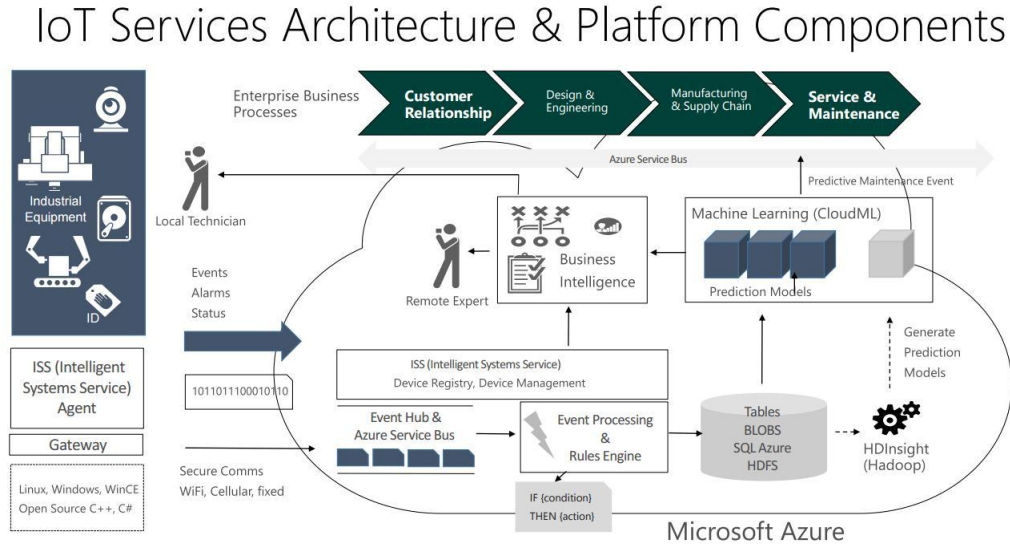


Figure 2.6: The graph shows a conceptual model of the global structure of a smart factory.

- Enabling monitoring conditions that provide manufacturers tools that increase OEE (Overall Equipment Effectiveness) performance considerably, hence plant performances.
- Revolutionizing product and service quality by determining which factors most and least impact.

Despite car makers worldwide seem to be a bit far from the actual application of digital concepts described above, FCA is moving toward the innovation in digital from lots of point of views. This master thesis was born thanks to the willingness of linking academical knowledge to plants' needs and trying to solve them with already present technologies and using the huge amount of data which automated systems produce. Before going in the details of WCM, following, the next section deals with a very quick overview on the classical structure of a car making plant, so that the reader can understand the practical context of the case study.

2.3 FCA plant structure

The car making process is hugely complex and difficult and requires a high degree of organization and the splitting of manufacturing processes that are described as follows ([17]):

- **Stamping** Production starts in the stamping shop, where gigantic presses transform rolls of metal into the main parts of the body of the car. The stamping shop is also equipped with a state-of-the-art metrology room where, before delivery to the body-in-white area, a pair of 3D camera robots use anti-reflective blue light to scan the components for imperfections.
- **Body-In-White** The next step in the process takes place in the body-in-white (BIW) area (shown in Figure 2.7), which is equipped with lot of robots. Assisted by sensors and cameras, these next-generation robots assemble the stamped metal parts to form completed bodies-in-white. A variety of joining techniques are used – including welding, gluing and screwing – and each process is subject to rigorous controls.



Figure 2.7: Assembling of the body metal components

- **Paint Shop** The assembled body is transferred from the BIW area to the paint shop, where it undergoes multiple washes to remove grease and other impurities. It is then submerged in the cataphoresis tank where, using an electro-chemical process, a uniform protective coating is applied to protect the vehicle from corrosive elements. Once dried, about 100 meters of sealant are applied to completely seal the body and prevent air and water entering the passenger compartment. The vehicle then enters a sterile chamber, where highly-efficient robots can paint the entire body in just 90 seconds – using only 3.5 kg of paint. Once the

paint has been baked, a team of specialists checks that the spraying has been carried out with perfect precision.

- **Assembly Shop** Finally, after the principal mechanical components (engine, suspension and transmissions) are fixed to the body with high-strength steel screws, wheels, brakes, seats, control panels, steering wheels, lights, the on-board entertainment and electrical wires and other systems are fitted as the vehicle moves down the assembly line (Figure 2.8) which is basically the field in which human made operations happen and transform the final product. Each component is sourced and delivered to the stations on the assembly line at precisely the right time. This complex logistics process requires methodical planning and organization supported by an advanced IT platform that connects factory, suppliers and logistics in real time. In addition, all workstations are equipped with eight anti-error devices that automatically stop the assembly line if an abnormality is detected, as well as terminals for logging completed activities and submitting suggestions for improvement. Members of a workstation team are trained on all tasks to enable rotation.



Figure 2.8: The assembly line of Melfi plant

In order to clarify the scenario of the plant in which there are the actors related to the Mura analysis, the following graph (Figure 2.9 summarizes how workers are organized within an FCA plant and the plant management system.

Assembly shop is composed by a lot of stations. Each station is the workplace of six operators performing some manual operations, and a Team Leader. Rather than the traditional top-down approach, FCA plants adopt a lean organizational model centered on production units, through direct and effective communication, the multi-disciplinary role of Team Leader ensures alignment with the

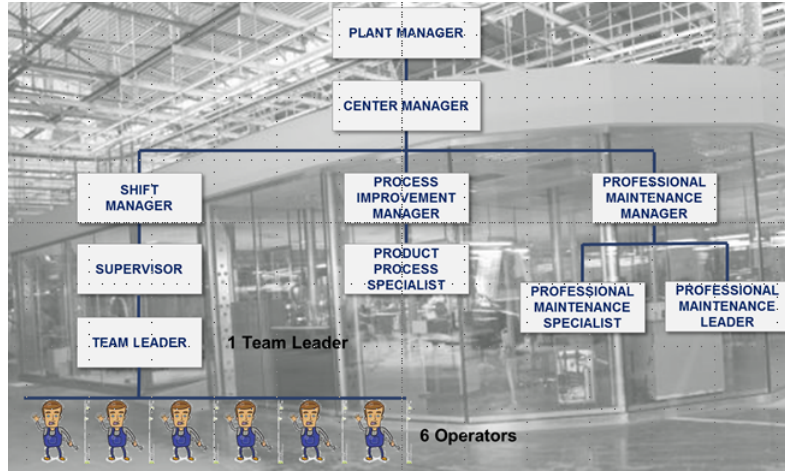


Figure 2.9: Global organization of an FCA plant.

highest production and quality standards. Therefore, the most involved part of the organization in the generation of Mura is represented by these figures, that are coordinated by the Supervisor, which manages the whole UTE (Unità Tecnologica Elementare), which is the set of all the stations.

Before analyzing how the work is accomplished, since the manufacturing is guided by the WCM principles as already said, the next section will focus on the methodology used by FCA, how it was born and how it is used nowadays.

2.4 World Class Manufacturing

2.4.1 Introduction to WCM

During 50s, Japanese car manufacturers started to develop several methods to optimize manufacturing processes since profit margins are usually very small: one of the most famous is the Toyota Production System (TPS). The aim of this methodology focused on eliminating all wastes from plants and gave birth to several new concepts such as Lean Manufacturing, Just In Time, Total Productive Maintenance and Total Quality Management [36].

World Class Manufacturing was born in the United States in the 80s. It is an innovation program based on the concept of continuous improvement which inherits all Japanese concepts mentioned above, but it is different from TPS since every kind of strategy is guided by the concept of "focusing" attentions and resources towards *attacking* every kind of waste and loss. This approach is lead by Cost Deployment (CD) pillar, which is the compass that highlights the main loss sources and

quantifies the benefits coming from eliminating them. It implies that each activity should be analyzed from its own economical impact on the plant balance. It is a working method to be applied every working day without an end point: this is the benefit brought by *continuous improvement* until reaching the World Class level and the condition of zero wastes and zero losses.

Every WCM team plant activity is oriented to the realization of projects called "Kaizen" which aim to reduce losses and eliminate their causes. The figure 2.10 shows the structure of the so-called *standard kaizen*, which is used for the analysis of the workflow of a standard problem-solving. The

Consulente WCM.com		STANDARD KAIZEN		N° Kaizen 1	
				Plant:	Italy
				Unit:	Warehouse Dept.
Theme: Implementation of Kan ban to the STOCK n°13023					
CATEGORY	TEAM	SAF	CI	FI	
LCS		RM	WO	PM	
		QC	LCS	EEM	
		PD	ENV		
		M.Pinfo	O.P.L	S.M.P.	
Problem Description with sketch		Probable Root Cause & Analysis		Possible Solution with sketch	
1. PLAN		2. DO			
Probable Root cause 1 Probable Root cause 2 Probable Root cause 3		Solution 1 Solution 2 Solution 3			
Ishikawa Diagram (4M)					
Standardization		Result		3. CHECK	
4. ACT		Result			
Are possible to use that Improvement in othe Area? If yes, Where?		Result			
Improvement autor		Kaizen open on:		Kaizen closed on:	
Verified by:		Cost		Benefit	
		Result		B/C	

Figure 2.10: Standard Kaizen: used for the implementation of a project aiming to the continuous improvement.

word "Kaizen" is the composition of two Japanese words: "*kai*" means changing, improvement and "*zen*" means better and refers to the step by step gain in order to reach a condition of zero accidents, zero losses, zero wastes, zero stock, maximizing the benefits-costs ratio and get the greatest satisfaction for both company and workers [2].

This methodology was introduced in Italy in 2005 by Fiat Group and contributed to the re-launch of the car manufacturer sector, thanks to Hajime Yamashina, Professor Emeritus at Kyoto Universality in Japan which played a key role, able to re-elaborate and contextualize the methodology in the European scenario. The greatest innovation introduced by the Japanese professor

deals the introduction of Total Industrial Engineering (TIT): *"A systyem of methods where the performance of labor is maximized by reducing Muri (unnatural operation), Mura (irregular operation) and Muda (non-value added operation), and then separating labor from machinery through the use of sensor techniques"*. TIT system integrated by professor Yamashina deals, hence, with the solution of manufacturing problems which realizes the continuous improvement by involving the whole operations staff through the usage of precise concepts as:

- Orientation to the whole system rather than the single department
- Inclusion of people in improving actions from the methodological point of view
- Knowledge of industrial engineering techniques
- Focus on people working in the plant

There are few revolutionary and distinctive aspects that differentiate WCM from the classical approaches [14]:

- Structured and strict approach organized in pillars and steps
- Strong attention to measurability
- Introduction to new topics (client service and people development)
- Structuring of elements such as planning, organization, leadership and motivation

All the group companies and suppliers joined to the program: Fiat, Maserati, Lancia, Alfa Romeo, Magneti Marelli, Teksid, Comau, CNH Industrial, and also Chrysler when is was acquired by the group which became FCA, i.e. Fiat Chrysler Automobiles. Today, more then 560 FCA plants apply WCM and the whole company gets benefits year by year, enough to export the application to other companies operating in totally different fields, offering training materials and coaching on World Class methodologies by following a business line lead by WCM Training and Consulting office. Technical aspects of WCM integrate perfectly with Industry 4.0 model, which forecasts the digitalization of production processes by applying some new enabling technologies (KET - Key Enabling Technologies), like 3D-print, advanced robotics, simulation systems, augmented and virtual reality, object communicating through the Internet of Things, new telecommunication paradigms and protocols and very powerful and cheap computation capabilities in order to reach as much as possible the status of "smart factory" [37]. All actions are evaluated by their capacity of affecting the processes performances, thanks to the actual evaluation of solutions' benefits in

order to guarantee sustainability of the industrial development of the company. At the moment, WCM methodology is implemented in every plant of the FCA Group, including historical ones like Mirafiori, Giambattista Vico in Pomigliano D'Arco, Melfi and Cassino.

2.4.2 WCM structure

WCM strategy is transverse to every process and involves all the company activities by using and implementing methods included in ten technical pillars and ten managerial pillars. This model deals with the determination of priority of action by means of the identification and analysis of wastes and losses in the production system called "Cost Deployment" pillar. As already explained, it is a kind of helm which guides actions toward the solution of the most onerous problems or conditions. So, the output of Cost Deployment is the identification of small plant area in which a problem has been identified, characterized and evaluated in terms of cost, called model area, where doing projects (Kaizen) and actions to face and attack the loss detected by Cost Deployment. The target is reached through implementing methods and tools included in the ten technical pillars and managerial pillars that ensure the management of production and operative problems and the commitment needed to carry on the program.

Managerial pillars are aimed to support the technical pillars and inform the management and the entire organization of WCM benefits. The main result of managerial pillars implementation is the allocation of resources and the commitment for the WCM program. They are still today in an early phase of both implementation and application. These pillars are part of the revolutionary concept lead by WCM, in which planning, organization and leadership, differently from Japanese model, play a very central role. Pillars are summarized in the following temple-like structure, shown in figure 2.11, where columns represent technical pillars, and basis represent the managerial ones. The whole continuous improvement process starts from the identification of a model area where pillars act in order to solve findings; then, whenever the result is reached, the project is expanded to the other plant areas and continues with the standardization and implementation of the found solutions in the expansion areas, until it arrives to the whole plant.

Each pillar is composed by seven different steps: each of which has an input, which is usually the problem to solve or an indicator to improve and an output, which is usually a performance indicator called Key Performance Indicator (KPI) and/or an activity indicator called Key Activity Indicator (KAI).

The verification and the achieving of different performances level of WCM plan is documented by a system of internal and external audit. They are used to validate the implementation of WCM

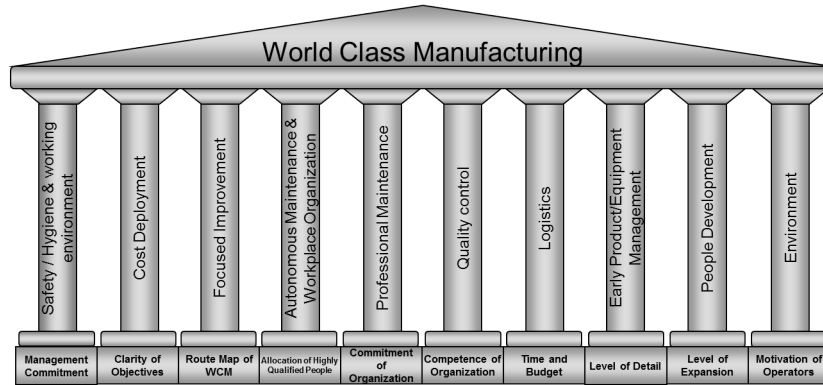


Figure 2.11: WCM temple structure: columns represent technical pillars. Basis represent managerial pillars.

toward the principles of continuous improvement. While internal audits are focused on the self-assessment and are conducted by own pillar leaders, external audits are assigned by World Class Manufacturing Association [19] and with the purpose of evaluating each of the twenty pillars and assign a score. The final score labels the plant level which can be Bronze, Silver, Gold or World Class.

The case study in question, is managed by and implemented within the "*Autonomous Activities*" pillar. Next paragraph focuses the attention on the explanation of the theoretical concepts behind it and how it is linked with this work.

2.4.3 Workplace Organization sub-pillar

After the explanation of the general structure of WCM, the focus moves into the pillar closely related to this work, which is "*Autonomous Activities*" pillar. It is actually divided into two sub-pillars: Autonomous Maintenance (AM) dealing with capital intensive areas and Workplace Organization (WO) dealing with labor intensive areas, that are areas requiring a high level of manually executed operations by operators. In FCA plants these areas are *assembly UTE (Unità Tecnologica Elementare)*. Since the aim of this master thesis is analyzing the variability of operation times made by human operators, the focus is pointed to the WO sub-pillar.

The main objective of WO pillar is increasing of productivity in labor intensive areas keeping the principle of Minimal Material Handling³ Furthermore, it must:

³Minimal Material Handling involves short-distance movement within the confines of a building or between a building and a transportation vehicle. In WO case, it deals with minimization people movements.

- Guarantee ergonomics and safety to operators
- Product quality through a robust process foolproof
- Respect production plans and realize high services to workers

Basic principles

There are some common principles that should help the comprehension of the discussion.

The first one is the concept of "action" that can be done into a plant and can be quantified from an economical point of view. Basically, there are different kind actions that can be performed within a production process:

- Value Added (VA) Action: is the time used to perform activities that actually transform and add value to the final product. E.g. Screwing a piece of body.
- Semi-Value Added (SVA) Action: is the time used to perform activities that are necessary to VA actions, but whose do not add a value to the product. E.g. Taking a tool.
- Non-Value Added (NVA) Action: is the time used to perform activities *useless and not requested by the process* that, therefore, do not add any value to the product. E.g. Walking, looking for a tool.

Another important principle is the Takt Time (TT). It is the production rhythm such that the market demand would be satisfied. It is calculated as follows:

$$TT = \frac{\text{Available time per day}}{\text{Clients demand per day}} \quad (2.1)$$

Takt Time is different from Time Cycle (TC), which is the time necessary to the completion of an assembly operation, and depends on the process. From the knowledge of both of them, is possible to get another important parameter, which is the number of operators (NoP):

$$NoP = \frac{TC}{TT} \quad (2.2)$$

Finally, the two most important concepts of an assembly line are shown in the figure 2.12 to better understand. The time cycle of the production line is determined by the time cycle of the bottleneck operation, that is the longest one in terms of time. The dissaturation is the difference between operation time cycle and the bottleneck operation time and reflects an amount of dead time in which the plant does not produce. The yellow part deals with all the actions related to

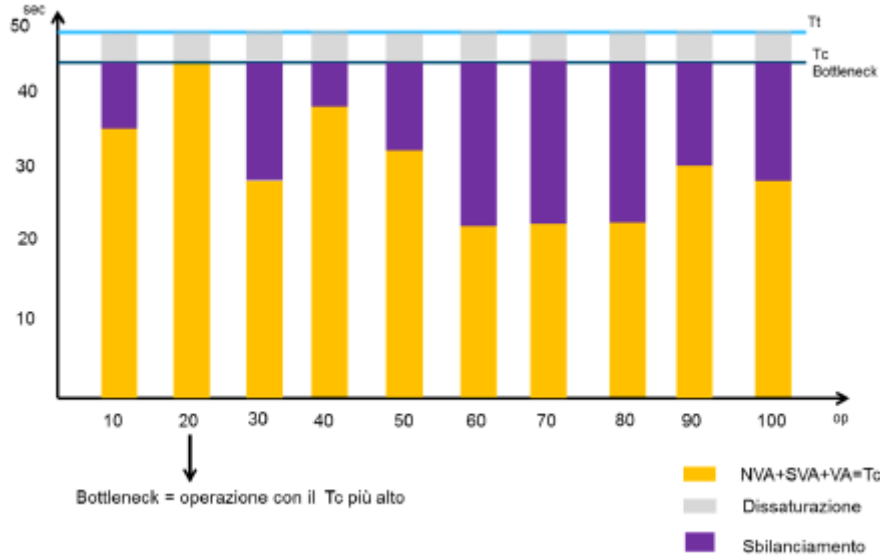


Figure 2.12: Figure shows 10 operators doing 10 operations each. The arrow highlights the bottleneck operations and "forces" the time cycle to be very close to the takt time.

the sequence of operations, while the violet part is the *unbalancing* and is the difference between the bottleneck operation and the time cycle of the single operation. This is a time in which some operators work, while other operators have already done their own operations. Hence, it is easy to understand that the more the small yellow towers are low, the faster is the production line, with very high consequences in the company business. The ideal situation would forecast that every operator performed the sequence of operations employing always the same time in order to reduce the time cycle, hence producing more cars, or to re-organize the unbalancing such that company earns on worker salary. Figure 2.13 summarizes a cost matrix made by Cost Deployment, which clusters losses per macro-fields, and is possible to notify that unbalancing is one of the main cost voices. Going deeper in the cost deployment analysis, figure 2.14 shows that the main pillar involved in the losses is WO in a stratification losses graph, and it is very higher than the others pillars. For this reason, is very important analyzing the main factors and the reasons for which this pillar is so critical. All these losses that involve WO are due to incorrect allocation of tasks among all assembly shop workers. The causes might be related to technical constraints, bad logistics organization or lack of capabilities by the blue collars themselves. Starting from the root-cause of the problem, different pillars are called to purpose to solve it.

Since the variability of the human-made operations is fully faced by the step 2 of Workplace Organization sub-pillar, for sake of clarity the step 1 should be briefly discussed in the next

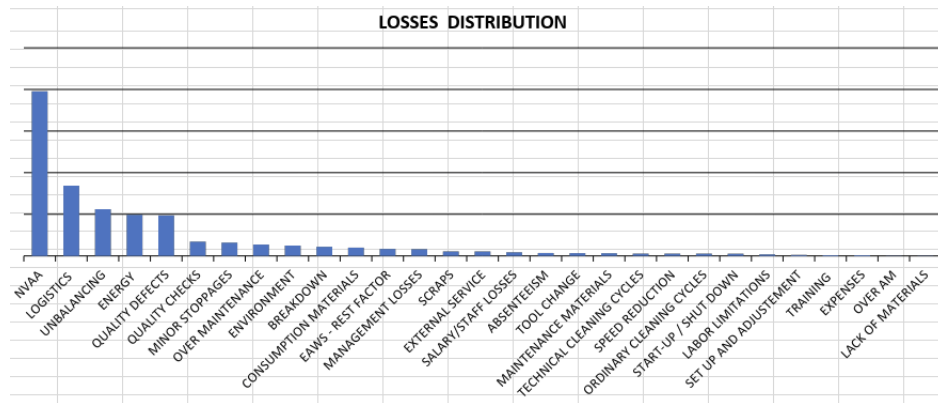


Figure 2.13: Actual C-Matrix from cost deployment. It shows as the main losses sources are NVAA, logistics and unbalancing; typical human related operations.

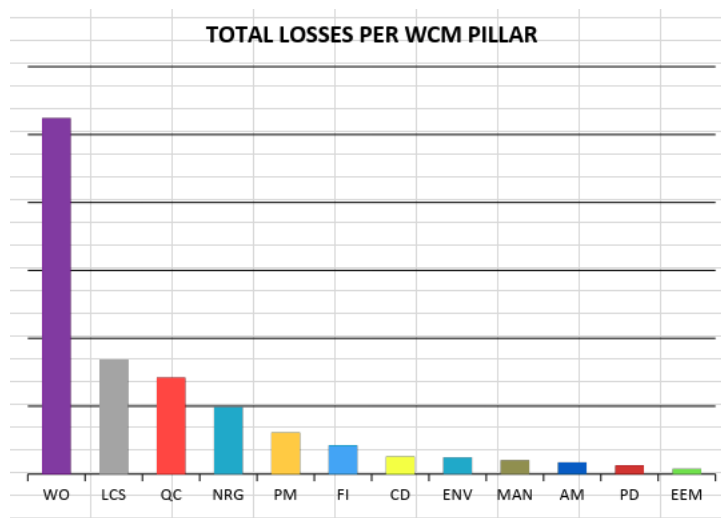


Figure 2.14: Total losses per WCM pillar. The first one is Workplace Organization, highlighting its importance in the plant business.

paragraph.

Step 1

The main activity of the first step of WO pillar is application of *5 S*, shown in Figure 2.15. It comes directly from the Japanese TPS [41], and has the purpose of getting the workstation *cleaned, ordered and organized* so as to keeping the process under control (with low variability) and standardizing tools positions and cleaning procedures. This methodology contributes in creation of a certain mindset able to keep the workplace cleaned and ordered getting easy the little continuous improvement actions. These operations are done in parallel with the first step of

"Logistics" pillar, where different operators come into contact and work with the same tools.

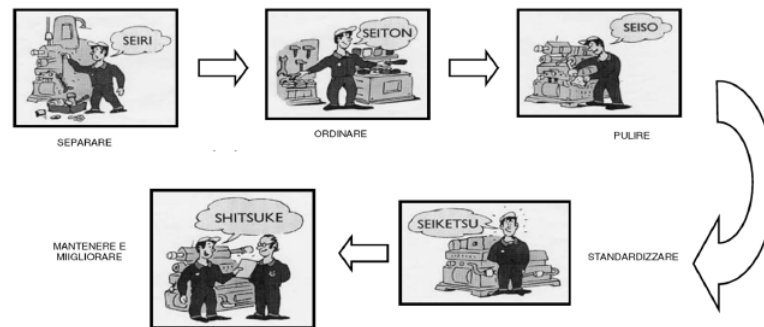


Figure 2.15: 5S application picture

Step 2

The step 2 of WO is the core of this master thesis. The aim of this step is the creation of a workplace in which it is easy to work, it is easy to work and the workers are efficient in total safety conditions, by attacking the main enemies of productivity, which are the so called *3 M*:

1. **Muri** means fatigue. Effects of Muri are speed reduction of activities, injuries, absenteeism caused by illness and dissatisfaction of the workers.
2. **Mura** means variability. Is the capstone of the wastes in industrial processes because it generates instabilities in the system with the consequent variation in the productivity and in the processes parameters.
3. **Muda** indicates the real wastes: Non-Value Added Actions. It represents all the activities that do not generate added value to the product or the service provided to the final client, hence it is useless cost to the company balance.

For sake of clarity, the figure 2.16 shows the composition of a Time Cycle for an operator. Is composed by VAA, SVA and NVAA. The red circle highlights the fact that waits (Muda) should also be taken into consideration for the line balancing.

Undoubtedly, whenever a problem of Workplace Organization must be faced, the 3Ms should be attacked fairly buttoned up, because, for example, is not reasonable thinking about deleting the variability without creating a safe and comfortable workplace, otherwise the fatigue would heavily condition the variability of the operation itself. Muri is solved by considering all operations requiring physical strength, postural fatigue and unpleasant operations. These concepts are summarized



Figure 2.16: The figure shows the composition of a Time Cycle for an operator.

by projecting the operation such that worker works in the so called golden zone [27], which is the most comfortable region of movement for a person.

Hence, whenever the process is stable, standardized and variability is reduced, is possible to attack Mudas, that are useless activities.

Since Mura is the core of this master thesis, the next paragraph goes into details highlighting the main definitions and the way in which is actually analyzed.

2.4.4 Mura

The concept of elimination of Mura comes, again, from the Japanese culture with its own TPS, thanks to the concept of "*Lean Manufacturing*" [3], in which the focus is upon improving the "flow" or smoothness of work, thereby steadily eliminating Mura through the system and not upon waste reduction itself.

Hence, Mura points out the fluctuations, variations, irregularities of workload. These factors involve the creation of some zones of the production cycle in which there is a serious overload that could create Muri, and zones in where there is an under load that could form wastes of time, hence Muda. The production flow, of course would suffer some troubles. So, investigating the causes of these fluctuations is very interesting, and it is the main scope of this work. Hence, Mura is the capstone of every waste (Muri and Muda) and that is why at the bottom of *lean production* there is the stability of the system that is obtained by eliminating the causes of the fluctuations

and standardizing activities. Moreover, a stable system is easier to be controlled and kept under control and does not require a lot of resources and it is safer than an unstable one.

But, starting from the assumption that a process has not Muri, measuring Mura is the actual challenge, because is not easy at all. Basically, the FCA approach is based on very difficult and expensive analysis. From a theoretical point of view, the ideal approach to face variability problems followed by FCA plants is cyclically repeated and is the following:

- Variation analysis of the time cycle of each operator/operation
- Reducing variation of the time cycle of the most critical operator, which is the one that has the highest variation level in operation times⁴
- Evaluation of the mean time cycle of each operator/operation
- Shifting the mean time of each operator/operation towards the best one⁵
- Now that operation is standardized, is possible to try to analyze the new standard and to work on the reduction of NVA

Whenever Mura is deleted for an operation, hence the process is totally stable and regular, the mean operation time is undoubtedly lowered because there is the certainty that the operation is so robust such that the variability would not affect the time cycle and so a line stoppage. The following graph, shown in the Figure 2.17 shows the shifting from a wide time distribution to a narrower one with a minor nominal time.

Whenever the step 2 of Workplace Organization pillar is done, there is a huge impact on the process. Considering again the example done in figure 2.12, is possible to explain conceptually how the unbalancing would change. The figure 2.18 compares the situation of the balancing before and after step 2 application. The most important concept shown in the graphs is, on one hand, the great reduction of Time Cycle that implies the growth of both dissaturation and unbalancing. On the other hand, is possible to notify that the most critical operations, before the application of step 2, was the second one. While after the step 2, the most critical set of operations, hence the one with the bottleneck time cycle, is the one labeled with 40.

As already explained, for sake of clarity, this process of standardization and stabilizing the process would continue until all operations last more or less the same time, such that there would be

⁴Attacking the operator with the highest variability gets the chance of correct an higher number of wrong actions.

⁵The operator that perform operations in the minor time will be also the best one, and its own standard will substitute the old standard in order to optimize the operation.

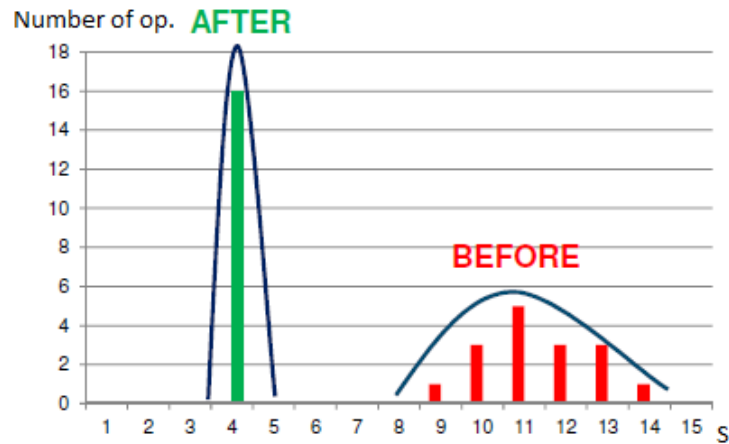


Figure 2.17: Effect on the nominal time of the results of a Mura analysis

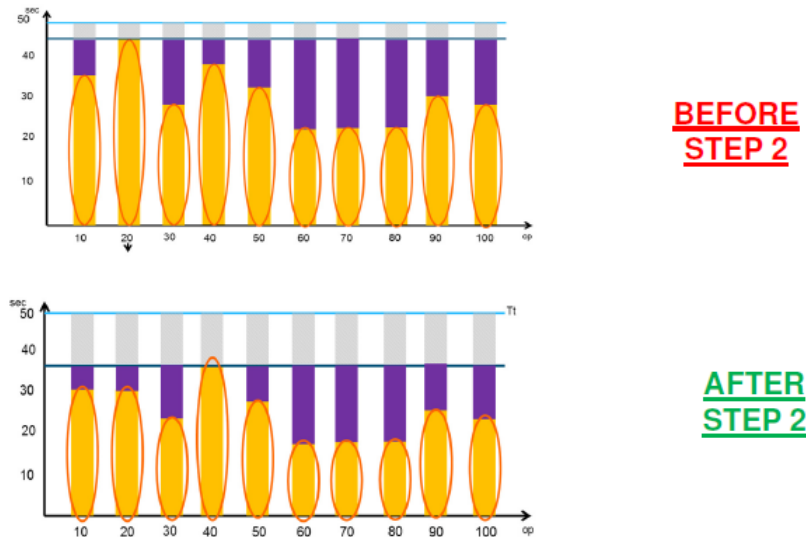


Figure 2.18: Re-organization of the labor intensive process due to application of Step 2 of WO pillar

a fair distribution of the workload, and the labor specialists would be able to re-organize properly the process, with all the benefits and advantages already discussed. The figure 2.19 shows the concepts already explained, but actually it happens during the implementation of step 3: thanks to the elimination of NVA, hence after the completion of the step 2 application, the Time Cycle is reduced. The economic benefit discussed above, coming from the re-organization of activities and the restoration of the balancing, such that a completely dissaturated operator could be assigned to other activities and line productivity is improved.

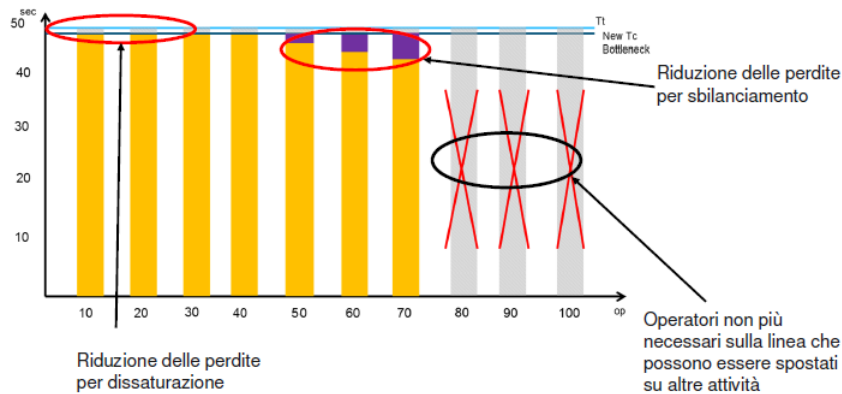


Figure 2.19: New line balancing coming from the elimination of NVA actions. Operators have been displaced to other tasks, with economic benefits for the plant.

In order to conclude the presentation of the pillar involved in this work, the last graph (shown in figure 2.20), describes a real situation in an FCA plant.

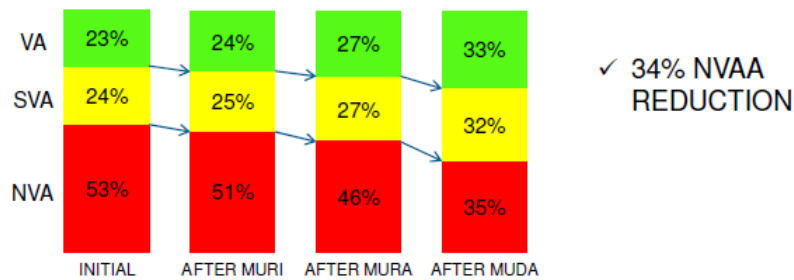


Figure 2.20: Actual Key Performance Indicator of the implementation of step 2 in a Fiat Power-Train plant.

The next chapter focuses more in details in the Mura analysis from the technical point of view. It explains how, at the moment, this problem is faced by people working in the plant and how data are gathered and analyzed.

Then, it will be explained in detail the core of this Master Thesis with the analysis conducted in collaboration with the WO pillar leader of Mirafiori plant, where some innovative data analysis tools are used, linking the discussed Industry 4.0 trend with traditional manufacturing methodologies.

Chapter 3

Analysis

3.1 Mura analysis: premise

Mura means loss due to irregular movements and not repetitive operations. The stakeholders involved in Mura analysis are basically Team Leaders, WO Specialists, Work Analysts (Industrial Engineers) and the higher levels of plant management. The aim is to identify variation in all man related operations, including semi-auto operations (machine interactions)¹.

At the moment, in FCA plants, the state of the art of Mura analysis is closely related to the "classical" approach, coming from Japanese methodology which forecasts the already mentioned *lean manufacturing*. Basically, analysis is based on observation of workers and the revision of the list of operations. Unless being observed for some time, this problem could be very difficult to be detected, or could not be recognized at all.

In a high quality assembly line, Mura covers a huge importance on process indicators. High quality reflects the fact that the Takt Time is considerably higher (about 6 minutes), meaning that the number of operations to be done on the body is huge with respect to a line in which the Takt Time is lower². Usually, the high quality production lines build the FCA premium brands, i.e. Alfa Romeo and Maserati.

¹Semi-auto operations refers to operations aimed to control robots status. Typically present in paint shop and body shop

²In the lines where the Takt Time is lower, typically, the mass market brands are produced, like Fiat and Lancia

First of all, is important to clarify that each operator has a precise list of operations³ to be done in a perfect order, from a theoretical point of view. Figure 3.1 shows an example of operation tag, with the sequence of actions be be made by an operator. Actually, this does not always happen because of a lot of factors, both human or external, e.g. related to the operation itself, or related to the availability of pieces or the tools to be used.

Postazione: WPA_FL1_077_SX_6300

Montaggio sedile anteriore sx

Saturazione / Tabella Descrizione Attività

N.B. La dissaturazione può essere annullata in qualsiasi momento dall'aggiunta di altre attività, da eventuali variazioni del Miv o da aumenti dei livelli produttivi.
- i dati riportati hanno carattere strettamente riservato e costituiscono segreto d'ufficio

N. Oper.	Modello	Caratterizzazioni	Tempo base	Prod.	Totale minuti
Descrizione operazione					
DA5193001001-4070	4070	4070-TUTTI I TIPI PROTEZIONE BATTICALCAGNO: Prendere da carteggio e posizionare sul batticalcagno anteriore sinistro a.1 l'utensile di protezione.	0,107	T5	6,002
DA7042101001-4070	4070	4070-TUTTI I TIPI SEDILE ANTERIORE SN: Prendere il pannello e ricolarlo con cinghia sul sedile, verificando assenza danneggiamenti su plastica o cavi. Eseguire la lettura della targhetta per verifica congruenza. Montare il sedile e disporlo in vettura.	0,771	T5	51,847
DA7042107001-4070	4070	4070-TUTTI I TIPI SEDILE ANTERIORE SN: Eseguire su NPI, conferma per esecuzione connessioni elettriche su sedile.			3,716
DA7042106001-4070	4070	4070-TUTTI I TIPI SEDILE ANTERIORE SN: Posizionare il sedile su vettura costruendo i due panni di riferimento. Asportare la cinghia dispendibile su carteggio. Depositare il pannello a lato.	0,36	T5	26,380
DA7042105001-4070	4070	4070-TUTTI I TIPI SEDILE ANTERIORE SN: Eseguire il collegamento di n.3 connettori con le rispettive derivazioni del sedile verificandone la tenuta.	0,302	T5	22,636
DA6234901030-4070	4070	4070-TUTTI I TIPI SEDILE ANTERIORE SX: Prendere ed imbastire n.2 viti set. e fissare con smittatore. Portare sedile tutto smontato. Prendere ed imbastire n.2 viti post. e fissare con smittatore. Portare sedile in posizione intermedia.	1,582	T5	110,640

Somma ponderata

Figure 3.1: Operations tag for the specified workplace, with each single operation's description

In order to improve the quality of the job of an operator, each single operation could have some further facilitation, especially for difficult or uncomfortable actions. This additive tool is called Job Element Sheet (JES). An example of JES is given in figure 3.2. The philosophy is that using very detailed work instructions, where everybody would be able to carry out the same tasks in exactly the same fashion, would reduce variability in executing the operation in question by favouring, on the other side, the standardization of the operation itself and the reduction of number of defects on the final product.

Therefore, having told about this premises, the next section deals with a brief discussion about the state of the art of Mura analysis in FCA plants in order to point out the weakness of the current methodology used and highlight the great advantages that would give an innovative approach based on Industry 4.0 principles and techniques.

³The list of operations that an operator should done within the TC is called *operations tag*




Job Element Sheet					Dept: Assembly shop	
AL PULASKI PLANT		Number JES:		Date:	Model:	
WORK OPERATION DESCRIPTION:				C.Op:		
ACTIVITY:	IMPACT:	Sketch	Phase	WHAT to do (sequence of work cycle)	HOW execute (key points)	
					WHY (consequences)	
			005	- PLACE BLOCK	- ON TOP OF THE TOOL - INSERT ON GUIDE AND PUSH UNTIL CLICK	- TO ENSURE THE LOCKING
			006	- PLACE WINDOW	- INSERTING THE LOWER FIN - INSERTING THE UPPER PIN	- TO ENSURE THE ALIGNMENT
			010			
			030	- INSERT REAR PINS	- PUSHING UP TO CLIP	- TO GUARANTEE THE FOLLOWING FIXING OF THE SEAL
			035			
			040	- PERFORM TIGHTENING	- UP EXTERNAL - DOWN INTERNAL	- TO AVOID WATER INFILTRATIONS

Figure 3.2: Job Element Sheet (JES): shows in a very detailed fashion the exact sequence of the operation to be performed. The example shows, also with a graphical aid for the sake of clarity, how to fix the window to the body of the door.

3.2 Mura analysis: state of the art

Mura is a very time consuming analysis that requires time studying at least thirty times video recording (thirty cycles), gathering operation times and statistical distribution analysis. This analysis is, hence, an observational activity that must be performed directly on the field and it could be very difficult to detect the more critical stations and, therefore, a criterion of prioritization.

The very first step for the detection of an area that requires a Mura analysis deals with a checklist. It is a tool, consisting in a series of simple questions, that helps to define the jobs that require video analysis.

Whenever a critical station or operation has been found, starts the analysis by element, that is the already mentioned recording of 30 cycles. This is done to all operators that do the same operation, every shift and every crew.

At the end of recording, is possible to build an histogram of operation times per each monitored operation executed. An example is shown in figure 3.3

Whenever the times histogram is built, data are analyzed by using simple statistical analysis. The two main parameters used are the following:

- **Statistical variance** gives a measure of how the data distributes itself about the mean of expected value. Unlike the range that only looks at the extremes, the variance looks at all the data points and then determines their distribution (figure 3.4).

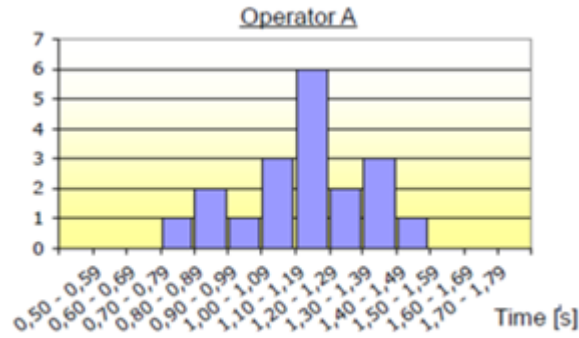


Figure 3.3: A result of the 30 cycle recording on a histogram for the "Operator A"

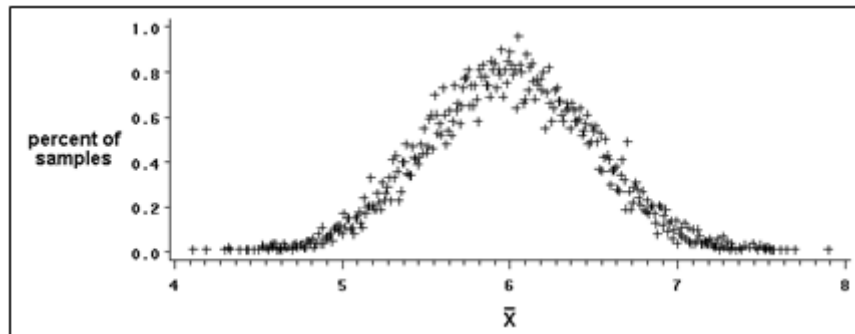


Figure 3.4: Distribution of operation times. In x axis, the mean time. In y axis the percentage of samples.

- **Variance** is a numerical value used to indicate how widely individuals of a population vary. If individual observations vary greatly from the group mean, the variance is big and vice versa.

Every single operation has a reference time, established from a theoretical point of view by labor specialists that quantify each single action during an operation. The target times are stored into a system called TiCon. The operation times statistical parameters are, then, compared with target times and is possible to detect how the operation times are displaced from the standard times.

Finally, the output of Mura analysis can be computed, starting from the discussed statistical parameters. In order to evaluate how the process is behaving with respect to the cycles, the Process Capability Indices (PCI) are used [40] [32] [1].

3.2.1 Weakness of the standard Mura analysis and advantages of Machine Learning

The current method has a lot of drawbacks. First of all, it uses standard tools which nowadays result to be very expensive and inefficient. Furthermore, the digitisation of manufacturing allows a great capability of generate and store data that are not efficiently used to perform these kind of analysis. The modern Machine Learning and Artificial Intelligence techniques would enable faster and smarter kind of computation and finding useful insights from data would result faster and easier [42].

Actually, ML is of course a very powerful resource that is widely used successfully in many fields. However, the field is very broad and even confusing which presents a challenge and a barrier hindering wide application. The aim of this handling points also out into the direction of demonstrating the real potential of ML applied to the manufacturing field in order to force the high management to innovate the traditional and standard approaches towards the continuous improvement.

The general advantages of ML have been established in previous sections stating that ML techniques are able to handle Non-Polynomial complete problems which often occur when it comes to optimization problems of intelligent manufacturing systems.

Another advantage of ML techniques is the increased usability of application of algorithms due to a lot of open sources programs and libraries, i.e. Tensorflow, Scikit-Learn [16] [30]. This allows (relatively) easy application in many cases and furthermore comfortable adjustment of parameters to increase the classification or whatever tasks performance.

The following paragraph explains the approach used to reach the thesis goals, highlighting the main aspects of Machine Learning and new Industry 4.0 tools applied to a real industrial case study.

3.3 Automatic Mura analysis

The basic idea of this work, is realizing a tool capable to automatize the complex operation of Mura analysis and detection, avoiding to carry out the expensive recording operations, and extending the analysis to the whole assembly line. With the introduction of new Information Technologies and tools, this task should be got in an easier way.

First of all, is important to clarify that operators should perform a list of operations. Some of

which are defined as *MES relevant*. As briefly explained in the introductory section, MES is an information system with the purpose of managing and controlling the productive function of the company. It has a direct link with the production line, with tools and operators and it acts as data collector of the whole plants system of FCA. Therefore, it covers a huge importance and it could be a great source of useful information.

With this tool, the plant will have a better work cycles control that will bring to an increasing time modelling accuracy with a production losses reduction expectations.

Monitoring user operations needs a comparison between effective and expected operation timing. Needed data are available on two different systems:

- plant MES that collects real time data from operations already executed.
- TiCon that stores nominal and expected process values.

Line users will be involved in the process. In particular, work cycles regarding screwing, traceability and quality reports will be monitored. The below picture (figure 3.5) shows the mainly features involved in the monitoring process at a single working station level. Whenever an operation has been fully completed, the system timestamps the event and record the entry in the database. In the upper part, are shown all the time delays related to the operations. Each of red highlighted sections, are referred to a single operation delay.

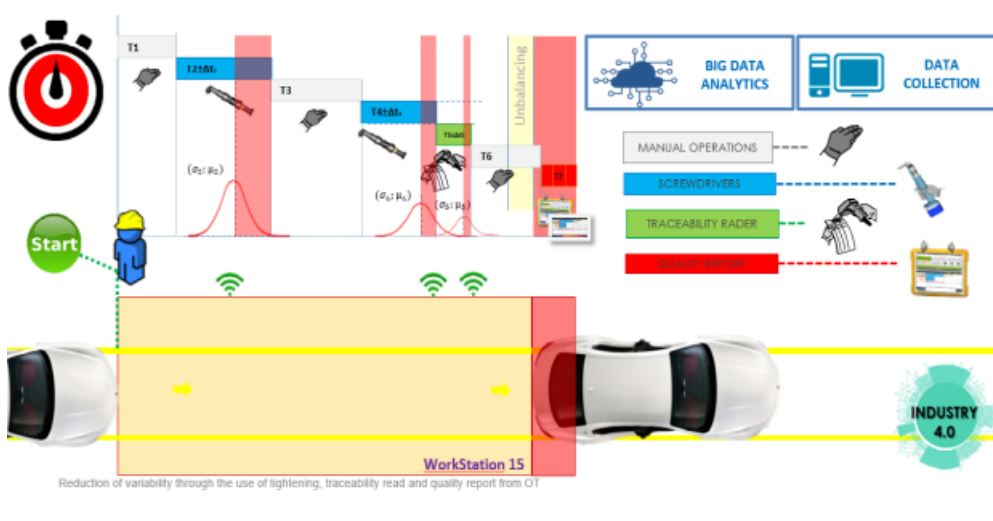


Figure 3.5: Example of process of a single workplace. Manuel operations are not considered in MES detection. In this case the line stops due to a delay coinciding with the last quality report operation (in red).

It is important to highlight that every operation time in MES is always the time period between

the T0 (start point) and the moment when the system receives the operation result from the line machines. Then, the instances are treated as traditional time series, therefore will be used some traditional tools for time series analysis [23].

From the operation times distributions, are evaluated the following parameters [26], which assume that population is normally distributed:

- Upper Specification Limit (USL) and Lower Specification Limit (LSL)
- C_p estimates a measure of how the process is distributed around a mean value.
- C_{pk} estimates a measure of how the process is distributed within the Upper and Lower Limits.

Both the indicators suppose that the distribution times were normally distributed. Actually, the state of the art analysis supposes that distribution times are normal.

Hence, the following analysis is basically composed by two parts: the first one is about the statistical analysis on operation times through the application of some statistical tools in order to check Gaussianity and to retrieve information about the statistical features of data. Then, the second one deals with contributors detection through the application of two ML algorithms.

3.3.1 Statistical Analysis

Statistics, independently from the field of research, represents an essential part of a study because, regardless of the study design, investigators need to summarize the collected information for interpretation and presentation to others [9]. The first step in a data analysis plan is to describe the data collected in the study. This can be done using certain data visualization techniques and generic descriptions [33] [24].

Before going into the detail of the analysis, is important to discuss briefly the data format to let the reader understand the kind of information that is possible to deal with. The figure 3.20 shows a sample of MES extraction from AGAP plant, where there are just two shifts, from 6 to 14 and from 14 to 22. Therefore, there are two teams alternating in the first and second shift. Basically there are data related to the operator and the operation time. There is also a label related to the line stoppage, and whenever it is present, the entry is "SI".

As already mentioned in the introduction section, the data source is represented by nine MES extraction corresponding to nine different operations. Basically, these extractions are contained into Excel files, each of which is imported into Python environment thanks to Pandas library,

3.3 Automatic Mura analysis

	C	D	E	F	G	H	K	L	M	N	O	P	Q
1	Operazione	Descrizione	Stato	Operatore	Workplace	Ingresso Workplace	Esecuzione	Uscita Workplace	Fermo Lit	Temp	Permaner	Percentua	Squadra
2	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:06:06.933	2017-12-01 06:08:38.197	2017-12-01 06:12:57.680	NULL	152	411.037	A		
3	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:12:57.693	2017-12-01 06:14:59.217	2017-12-01 06:19:51.730	NULL	122	414.029	A		
4	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:19:51.740	2017-12-01 06:21:53.020	2017-12-01 06:25:52.460	NULL	122	361.034	A		
5	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:25:52.477	2017-12-01 06:28:05.350	2017-12-01 06:31:27.450	NULL	133	335.040	A		
6	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:31:27.457	2017-12-01 06:33:18.913	2017-12-01 06:34:58.940	NULL	111	331.034	A		
7	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:36:58.957	2017-12-01 06:39:21.143	2017-12-01 06:42:29.373	NULL	143	331.043	A		
8	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:42:29.370	2017-12-01 06:45:03.263	2017-12-01 06:47:54.863	NULL	154	325.047	A		
9	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:47:54.870	2017-12-01 06:49:54.327	2017-12-01 06:53:05.327	NULL	120	311.039	A		
10	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:53:05.337	2017-12-01 06:55:11.233	2017-12-01 06:59:34.970	NULL	126	389.032	A		
11	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 06:59:34.983	2017-12-01 07:01:36.287	2017-12-01 07:06:05.983	NULL	122	391.031	A		
12	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:06:06.000	2017-12-01 07:08:15.740	2017-12-01 07:11:31.620	NULL	129	325.040	A		
13	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:11:31.637	2017-12-01 07:13:59.533	2017-12-01 07:17:39.743	NULL	148	348.040	A		
14	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:17:39.757	2017-12-01 07:19:39.763	2017-12-01 07:23:55.713	NULL	120	376.032	A		
15	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:23:55.730	2017-12-01 07:26:11.113	2017-12-01 07:29:18.500	NULL	136	323.042	A		
16	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:29:18.517	2017-12-01 07:31:52.787	2017-12-01 07:35:20.760	NULL	154	362.043	A		
17	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:35:20.757	2017-12-01 07:37:28.117	2017-12-01 07:41:39.693	NULL	128	379.034	A		
18	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:41:39.707	2017-12-01 07:43:35.183	2017-12-01 07:47:31.120	NULL	116	352.033	A		
19	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:47:31.140	2017-12-01 07:49:26.853	2017-12-01 07:53:50.310	NULL	115	379.030	A		
20	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 07:53:50.323	2017-12-01 07:55:51.940	2017-12-01 08:00:06.377	NULL	121	376.032	A		
21	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:16:44.640	2017-12-01 08:18:40.097	2017-12-01 08:22:28.657	NULL	116	344.034	A		
22	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:22:28.670	2017-12-01 08:24:31.737	2017-12-01 08:28:41.527	NULL	123	373.033	A		
23	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:28:41.540	2017-12-01 08:30:57.980	2017-12-01 08:34:25.340	NULL	136	344.040	A		
24	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:34:25.350	2017-12-01 08:36:19.913	2017-12-01 08:40:54.677	NULL	114	389.029	A		
25	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:40:54.693	2017-12-01 08:43:57.077	2017-12-01 08:46:38.660	NULL	183	344.053	A		
26	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:46:38.677	2017-12-01 08:48:36.487	2017-12-01 08:52:39.443	NULL	118	361.033	A		
27	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:52:39.460	2017-12-01 08:54:49.677	2017-12-01 08:58:30.933	NULL	130	351.037	A		
28	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 08:58:30.940	2017-12-01 09:00:52.450	2017-12-01 09:04:20.700	NULL	142	350.041	A		
29	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 09:04:20.717	2017-12-01 09:06:21.213	2017-12-01 09:10:25.780	NULL	121	365.033	A		
30	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 09:16:46.267	2017-12-01 09:18:35.627	2017-12-01 09:22:47.910	NULL	109	361.030	A		
31	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 09:22:47.927	2017-12-01 09:25:59.960	2017-12-01 09:28:53.203	NULL	192	366.052	A		
32	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 09:28:53.197	2017-12-01 09:31:06.063	2017-12-01 09:34:44.750	NULL	133	351.038	A		
33	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 09:34:44.740	2017-12-01 09:36:43.690	2017-12-01 09:42:01.727	NULL	119	437.027	A		
34	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 09:42:01.773	2017-12-01 09:43:44.477	2017-12-01 09:48:05.683	NULL	103	364.028	A		
35	SCR@A01@03A_000>SCR - tr1_03a fx tubi freno	OK	Op_1	WPA_TLO_003_DK	2017-12-01 10:05:45.593	2017-12-01 10:08:38.207	2017-12-01 10:12:50.413	NULL	173	425.041	A		

Figure 3.6: Sample of MES extraction from AGAP plant.

which is an open source library for data structure, manipulation and visualization [28].

The data structure has been organized such that every operation is inserted into a list, and each element of the list is a dictionary containing a lot of information like the dataframe itself, statistical parameters and other useful information keeping easy the readability and the management of accessing data (listing 3.1).

Listing 3.1: Data opening and creation of the data structure

```
import pandas as pd
import glob, os

def getOperationsList(path):
    os.chdir(path)
    operations = []
    filenames = []
    # Loop over EVERY xlsx file and appends dataframe to operations list
    for file in glob.glob("*.xlsx"):
        df = pd.read_excel(file, "DATI")
        filenames.append(file)
        dataframes = {}
        dataframes["OperationsName"] = file
        dataframes["Dataframe"] = df
        operations.append(dataframes)
```

```
return operations
```

Before applying any kind of analysis or machine learning algorithm, the dataset must be cleaned through proper data cleaning analysis and outliers detection.

Data cleaning

In order to perform the statistical analysis, the very first phase to be done is data cleaning. Since the aim of this work is analysis of variability of human made operations, is needed taking into account just operations without line stoppages, because they bring a lot of variability in operation times. Hence, the listing 3.2 aims to delete all the entries which cause a line stoppage and every non-valid entry, like zeros or NaN columns.

Listing 3.2: Code showing the very first step of datasets cleaning

```
import numpy as np

cols = ["CIS", "Sequenza", "Operazione", "Descrizione", "Stato",
        "Workplace", "Inizio_Fermo_Linea", "Fine_Fermo_Linea",
        "Uscita_Workplace", "Fermo_Linea", "Percentuale", "Ingresso_Workplace",
        "Permanenza"]

def getCleanedOperationTimes(dictionary, taktTime):
    "—>NOTE: the operation time should be in the LAST column"
    df = dictionary["Dataframe"]
    df = df[df.Tempo < taktTime]
    columns = df.columns
    appoggio = df["Fermo_Linea"]
    appoggio = appoggio.dropna(axis=0)
    indicidaeliminare = appoggio.index.values
    df = df.drop(df.index[indicidaeliminare])
    df = df.drop(cols, axis=1)
    dictionary["Dataframe"] = df
```

Since the first step of cleaning is done, is possible to discover the time distribution' features without performing any filtering action. Is important to say that in the dataset there are both team A and team B, therefore from now, every kind of visualization will be made for both teams.

Scatter plot Is a graph where two variables of a certain dataset are reported on a cartesian coordinate system. Data are visualized through a collection of points each of which with a position on horizontal axis determined by a variable and on vertical one determined by the other variable.

Box plot Is a method for graphically depicting groups of numerical data through their quartiles⁴. The central box represents the values between the first (Q_1) and the third (Q_3) quartile. The difference between the third and the first quartile gives the Inter-Quartile Range (IQR). The horizontal red line represents the median value, which is the central value of the distribution. In the end, the lines correspond to the lower ($Q_1 - 1.5 \times IQR$) and the upper ($Q_3 + 1.5 \times IQR$) values. Outliers may be plotted as individual points. Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The figure 3.7 shows an example of box plot for a Gaussian distribution population⁵.

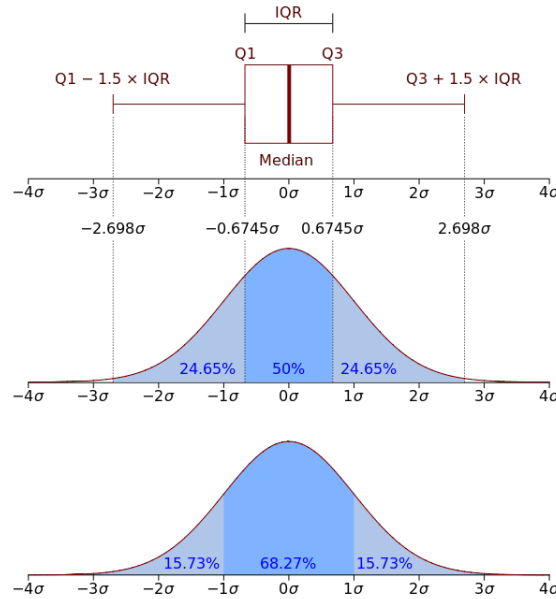


Figure 3.7: Box plot example for a Gaussian distribution.

Histogram Is a kind of bar graph, in which continuous or discrete data are divided into "bins", in x axis, representing the range of the variable and then counting how many values fall into each

⁴Quartiles are values that split a statistical population in four equal parts.

⁵Example written in the Wikipedia web page: https://en.wikipedia.org/wiki/Box_plot

interval, in the y axis. The represented histograms are normalized, meaning that each bin is divided by the total number of entries, showing the relative frequencies.

For sake of clarity, only a few of significant operations are taking into account for this analysis. The Takt Time of the line is about six minutes: so, every operation time lasting more then this interval are not considered in the visualization plots (method shown in 3.3).

Data are also extracted in a period in which the saturation remains still the same, meaning that operations procedures are the same and do not change. It is important to clarify it because sometimes plants use to re-balance workload, and so operations' sequence may change.

Listing 3.3: Method that plots scatter plot box plot and times histograms for every team presents in the dataframe

```
import matplotlib.pyplot as plt
from scipy.stats import lognorm
import matplotlib.dates as mdates

def plotTeamOperationTimes(dictionary):
    df = dictionary["Dataframe"]
    opName = dictionary["OperationsName"]
    count = 0
    f, ax = plt.subplots(nrows=len(df.Squadra.unique()), ncols=3)
    for sq in df.Squadra.unique():
        df_sq = df.loc[df.Squadra == sq]
        ax[count, 2].hist(df_sq.Tempo, bins=80, normed=True,
            label="Times□histogram")
        ax[count, 2].grid()
        ax[count, 2].legend()
        ax[count, 2].set_xlabel("s")
        ax[count, 2].set_title("Operation□times□frequencies□Team□" + sq)
        ax[count, 1].boxplot(df_sq.Tempo, 0, '.')
        ax[count, 1].grid()
        ax[count, 1].set_ylabel("s")
        ax[count, 1].set_title("Boxplot□Team□" + sq)
```



```
ax[count, 1].legend()
df_sq["Esecuzione"] = pd.to_datetime(df_sq["Esecuzione"])
ax[count, 0].set_xticklabels(df_sq["Esecuzione"].values, rotation=35)
ax[count, 0].plot(df_sq["Esecuzione"].values, df_sq.Tempo, '.')
myFmt = mdates.DateFormatter('%m%d')
ax[count, 0].xaxis.set_major_formatter(myFmt)
#ax[count, 0].rot
ax[count, 0].grid()
ax[count, 0].axhline(y=np.median(df_sq["Tempo"].values),
                    color='r', linestyle='-', label="Median_time")
ax[count, 0].set_title("Scatter_plot_Team_" + sq)
ax[count, 0].legend()
ax[count, 0].set_ylabel("s")
count = count + 1

plt.legend()
plt.suptitle(opName)
plt.rcParams["figure.figsize"] = (15,10)
plt.show()
```

The figure 3.8 shows the first operation for a certain workplace, divided into the two teams that perform the operation: The very first comment to be done by analyzing the graph is that the operation is performed more or less coherently. Median times are very similar. On the other side, the boxplot shows a huge number of outliers due to operation times very higher then the median time. In the end, the histogram shows that distribution is not Gaussian at all, but it seems to be lognormal.

The second operation which is interesting to be discussed is showed in figure 3.9. Again, the distribution of times seems to be lognormal. In this case, there is a bigger difference between the teams. The first one seem to work more standard than the second, concerning the scatter plot and the histogram. Operation times of B team are distributed along the whole time cycle, even beyond the limit of six minutes. This is also confirmed by box plot where, the A team has a lot of outliers in the upper part, and the second one do not seem to have outliers. This happens because the population is so distributed enough to cover all the Time Cycle. Hence, this operation should

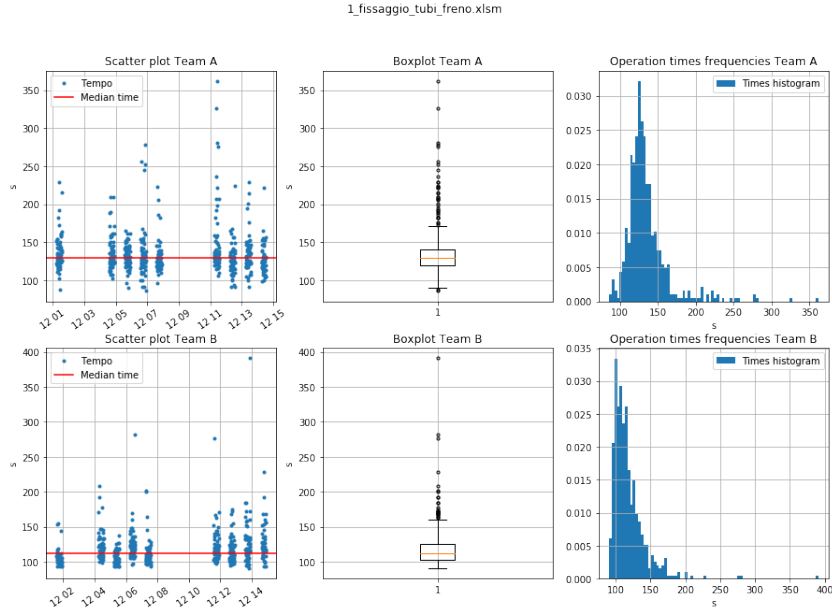


Figure 3.8: Visualization plot operation 2: scatter plot, box plot and times histogram

be investigated more in depth, especially for what concerns the B team.

This is a perfect example of how heavily the premium brand assembly line impacts the variability of operations.

For what concerns the last figure 3.10, it shows an interesting pattern in the time histogram. Distribution appears to be as a bimodal distribution: by definition it is a continuous probability distribution with two different peaks (local maxima) in the function.

This behaviour reflects the presence of two different prevailing mean operation time as consequence of the presence of two different operators per each team, or two very different kind of customization that cause a huge difference in timing.

Normality tests

After the removing of the main outliers with the explained method, the next step of pipeline experimented forecasts the application of two statistical tests capable to "how" the distribution of times is Gaussian, because, as already mentioned, the parameters used to control process capability assume the normality of population: when the process has a non-normal distribution, classical PCIs will be inappropriate and can misled the assessment of process capability[34].

Both tests start from the so-called null hypothesis, which is a statement about the probability

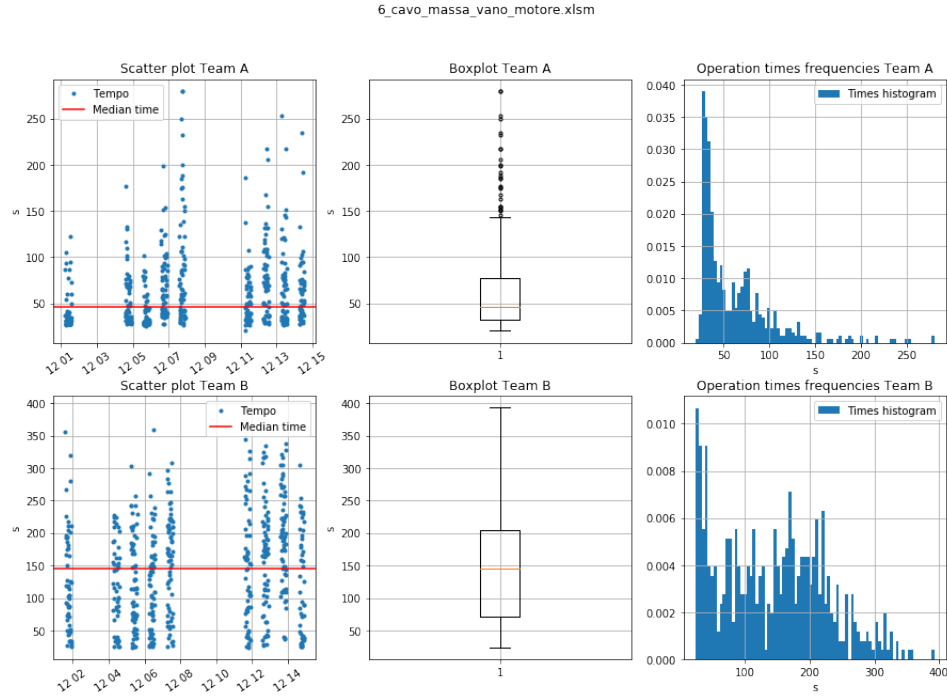


Figure 3.9: Visualization plot operation 6: scatter plot, box plot and times histogram

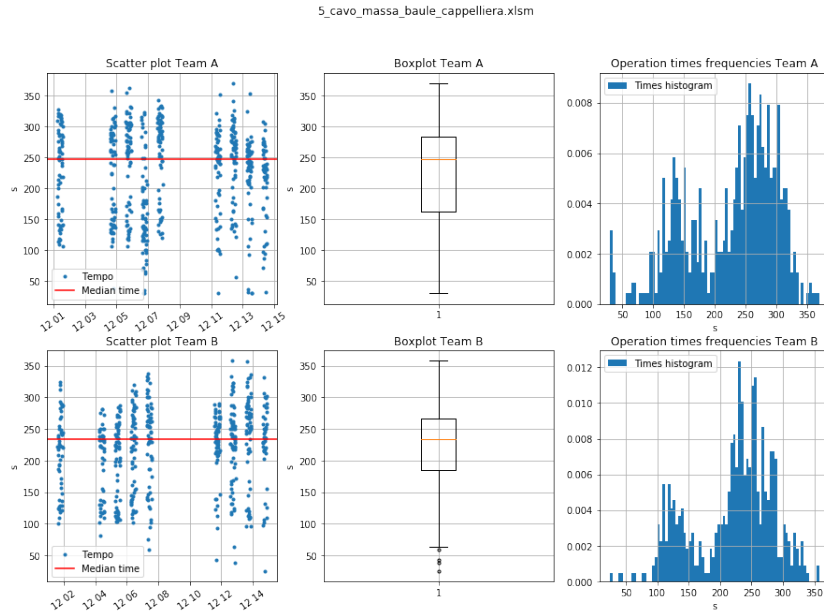


Figure 3.10: Visualization plot operation 5: scatter plot, box plot and times histogram

density function, which is, in this case that time distributions are normal. The hypothesis is testable

on the basis of observing the process that is modeled via a set of random variables, comparing the two distributions. The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability, called the significance level. Typically, significance level is a threshold set to 0.05. If the test's result is lower than the significance level, claiming that distribution is not normal is impossible. Otherwise, data are statistically significant such that it is possible to state that they do not belong to a normal distribution, hence null hypothesis could be rejected. In this case, it would be impossible to use the process indicators actually used now.

According to the available literature, visual analysis of a distribution could be powerful and reliable as well.

Shapiro-Wilk test It is a non-parametric test⁶. The null-hypothesis of this test is that samples come from a normally distributed population. The significance level of the test is defined as 5%. Hence, if the p-value of the test is lower than that value, then the null hypothesis is rejected and there is not proof that tested data come from a normally distributed population. On the other side, if the p-value is higher, the null hypothesis can not be rejected and the test can not state anything else. Said that, the test has some weaknesses: for example, if the sample dimension is large, the test may not sense significant effects of eventual relevant data [15].

D'Agostino-Pearson test Is is a non-parametric test. It has the same level of significance of the previous one (5%). This test aims to prove the null hypothesis that a sample comes from a normal distribution. It is based on D'Agostino and Pearson's test that combines skew and kurtosis⁷ to produce an omnibus test of normality. This test is useful for moderate size datasets.

The following graph (Figure 3.11) shows that almost every operation fails the normality test. Entries are divided into teams, because each team has its own operator working on that operation and it is independent from the others. If the cell is flagged as "False", hence the null-hypothesis can not be rejected.

After having demonstrated that the most of distribution is not Gaussian, the discussion goes on in the direction of unpacking data, almost feature by feature and find the variation patterns in

⁶Non-parametric test means that statistics is distribution-free, hence there are not assumptions on distributions' parameters.

⁷Kurtosis is a displacement from the statistical normality.

Operator	Operation	Gaussian Shapiro	Gaussian DAgostino
Team A	ssaggio_centralina_pompa	false	false
Team B	ssaggio_centralina_pompa	false	false
Team A	vo_massa_vano_motore.	false	false
Team B	vo_massa_vano_motore.	false	false
Team B	ssaggio_centralina_cba.x	false	false
Team A	ssaggio_centralina_cba.x	false	false
Team B	raccio_oscillante_sup.mo	true	true
Team A	raccio_oscillante_sup.mo	false	false
Team A	fissaggio_tubi_freno.xlsr	false	false
Team B	fissaggio_tubi_freno.xlsr	false	false
Team A	ssa_sx_fianc_ant_post.li	false	false
Team B	ssa_sx_fianc_ant_post.li	false	false

Figure 3.11: Summary of normality test for the available operations.

operation times, in order to allow the final user to choose certain criteria of more in depth visualization criteria and to investigate the possibility of find a way to extract a performance indicator for the process.

Summing up the normality tests part, only the 11.1% of the available operations do not fail the statistical tests. It means that process has different problems in terms of stability, hence, before using process indicators that assume the normally distributed operation times, it is *needed* to act over the process structure and find some ways to keep it stable, by using WCM methodologies, i.e. training sessions to the workers using One Point Lessons (OPL) or improving the JES frequency on assembly line.

It is unrealistic to expect that data visualization tools and techniques will unleash a barrage of ready-made stories from datasets (figure 3.12). There are no rules, no ‘protocol’ that will guarantee the certainty of finding useful information by just visualizing data. But undoubtedly, it makes sense thinking about finding some input to direct the investigation towards certain directions and methodologies [24].

With the insights gathered from the last visualization is possible to have an idea of what to

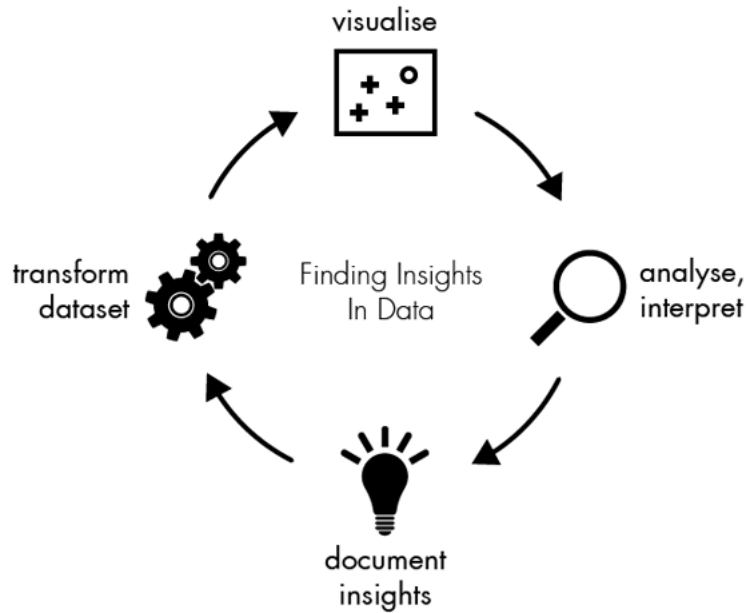


Figure 3.12: Conceptual schema for data visualization [24]

do next. It allows the detection of some interesting patterns in the dataset which is possible to investigate more in detail by transforming and manipulating data:

- **Zooming**: to have look at a certain detail in the visualization aggregation to combine many data points into single or more groups.
- **Filtering**: to temporarily remove data points that are not major focus.
- **Outliers removal**: to get rid of points that are not representative for the current dataset.

Outlier detection

As shown in the very first step of the visualization, there are different outliers from the box plot view. But this technique of outliers detection works well for normally distributed variables because it supposes that population is fairly distributed [10]: hence, outliers may be evidence of a contaminated data set; they may be evidence that a population has a non-normal distribution; or, they may appear in a sample from a normally distributed population. The following method 3.4 has been used to flag outliers with 1 and non-outliers with 0.

Two categories of outlier are defined, in general: "Additive Outliers (AO)" where a single point, or a group of points is affected and "Innovative Outliers (IO)" where an innovation to the process affects both an observation and the subsequent series [38]. In this case outliers are samples of AO

because, from a theoretical point of view, the outlier should be isolated itself and not conditioning the following process.

Listing 3.4: Method used to add an "outlier" column and flag the entries as outlier or not.

```
def outliersDetection(dic):
    df = dic["Dataframe"]
    #df_a = df.loc[df.Squadra == "A"]
    qua_1, qua_3 = np.percentile(df.loc[df.Squadra == "A"].Tempo.values, [25, 75])
    iqr = qua_3 - qua_1
    lb = qua_1 - (iqr * 1.5)
    ub = qua_3 + (iqr * 1.5)
    qua_1, qua_3 = np.percentile(df.loc[df.Squadra == "B"].Tempo.values, [25, 75])
    iqr = qua_3 - qua_1
    lb_b = qua_1 - (iqr * 1.5)
    ub_b = qua_3 + (iqr * 1.5)

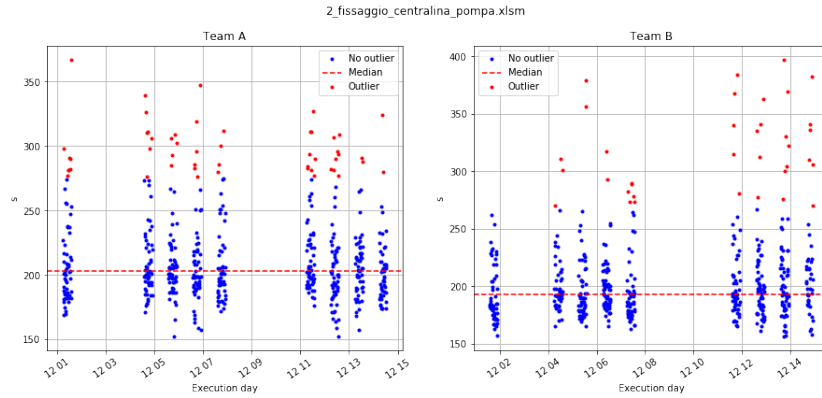
    df["Outliers"] = np.where((((df.loc[df.Squadra == "A"]["Tempo"] < lb) |
                                (df.loc[df.Squadra == "A"]["Tempo"] > ub) |
                                (df.loc[df.Squadra == "B"]["Tempo"] < lb_b) |
                                (df.loc[df.Squadra == "B"]["Tempo"] > ub_b))),
                                1, 0)

    dic["Dataframe"] = df
    #return df
```

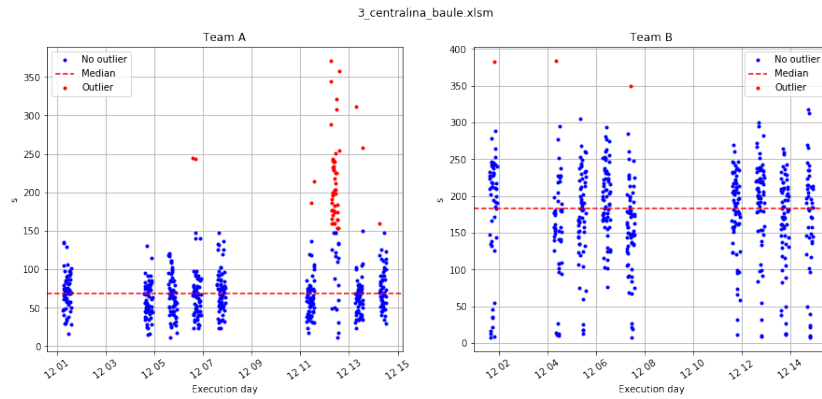
The effect on this flag operation is shown in the following scatter plots (figure 3.13).

Since the operation number 3 has a strange behaviour during a single day of production, the next step of the unpacking analysis is plotting a box plot per every execution day in order to check in detail if and how the execution day act in the global median time evaluation. Figure 3.14 refers to a severe variability problem in that station because the median operation time is three times higher than what happened the other days, highlighting a low number of outliers for the single production days. Hence is necessary to deeper in analysis of outliers in order to understand why the variability is that increased.

Figure 3.15 shows the said operation clustered by the execution day and worker and basically there are two situations appearing useful. The first one is that the operator named as "Silvia" is present into the dataframe but actually does not perform any operation. It could be due to a



(a) Operation 2 made more or less coherently by both teams.



(b) Operation 3 made in a widely different way by both teams. In Team A is possible to individuate a very critical day in terms of outliers.

Figure 3.13: Scatter plots for two operations: outliers are labeled by red dots.

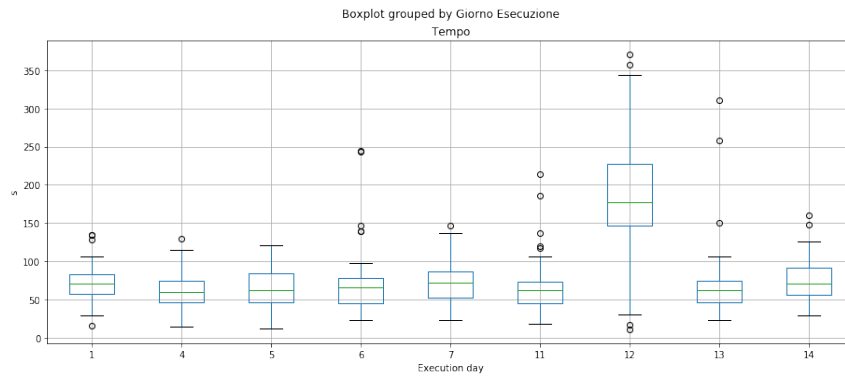


Figure 3.14: Box plot per each execution day for operation 3.

problem in line, and the operator changed station momentarily, hence it could be considered as an outlier. The second observation is more interesting. Is possible to notify that the operation is

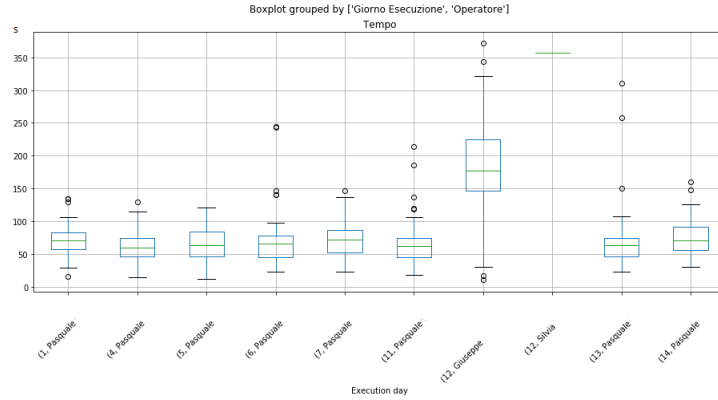


Figure 3.15: Box plot of operation clustered per execution day.

always made by the same operator called "Pasquale", while the twelfth of December the worker is changed and the new one has introduced a huge level of variability (the relative box plot covers almost the whole time cycle).

So, this simple observation could suggest that the operation is not properly made by the operator and consequently it introduces variability. It can also suggest to improve the mechanism of operation tag or the already mentioned JES.

Going deeper in the analysis of the single operator, the next step aims to investigate what happens in the most critical day explored above. It could be interesting because some correlations could highlight moments of the day more critical than others like the early phase of the shift in the morning, of the last phase of the shift, when maybe the operator is going to be tired. Figure shows the trend of operation times for the said day.

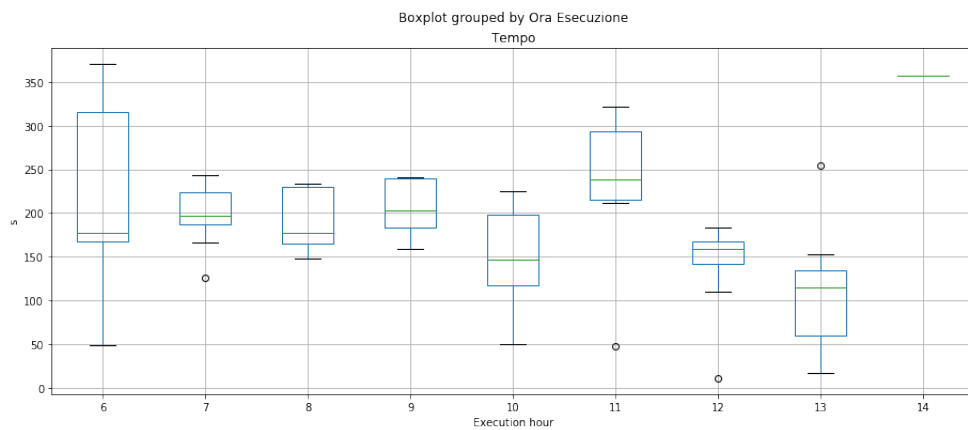


Figure 3.16: Box plot of operation clustered per hour of the day. It shows a huge variability during the early phase of the shift.

The graph mentioned (figure 3.16) shows a very predictable pattern: at 6 AM, hence when the shift starts, variability is very higher than the other hours. By the way, even median times are definitely far from being constant over the considered time.

This is an example of how the tool would be used by WO analysts: it could be a very first analysis to guide the attention of stakeholders towards the actual problems that cause Mura. Unfortunately, the available data is, until the realization of this Master Thesis, very reduced. Therefore, the statistical analysis is not that reliable because data agglomerate a lot of different scenarios, cases and also different operators.

Moreover, the effect of car customization has not been mentioned so far. An high quality assembly line, for FCA company, reflects the high level of complexity and quality of the final product. This implies that there is also a huge number of possible customization: in fact, every body crossing the line has its own customization (optional) code, called "*mix*", e.g. right-hand drive, particular internal materials, etc. Unfortunately, the datasets used have not information about optionals, hence, data is pretty contaminated by the presence of operations with different features.

The following example shows the utility of the tool for more "strange" distribution times, like the case of operation 5 (figure 3.10). In this case the times' distribution is bimodal, highlighting the presence of two different events that characterize the distribution without influencing each other. Again, without going into execution day or hour level of detail, the tool shows that the main variability component is related to human worker (figure 3.17)

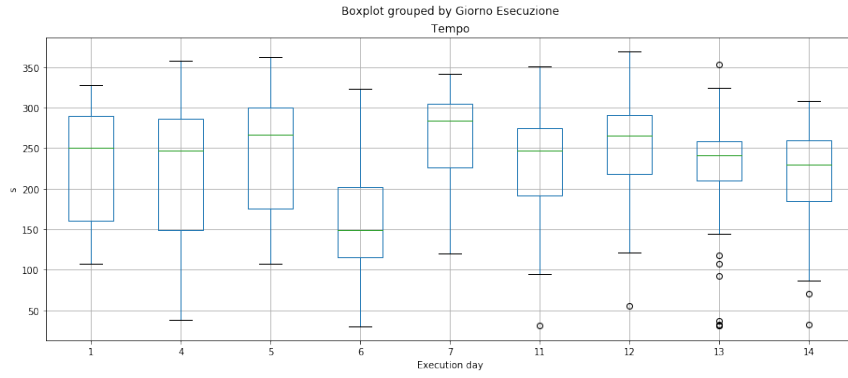
Again, team B seems to work a little bit better than the other one. In team A there are excessive fluctuations of median times: it could reflect the presence of different operators, working differently each other. Through visualizing unpacked data for what concerns team A, there will be another prove that operator introduces a huge degree of variability in operation times (figure 3.18).

Ordered operation

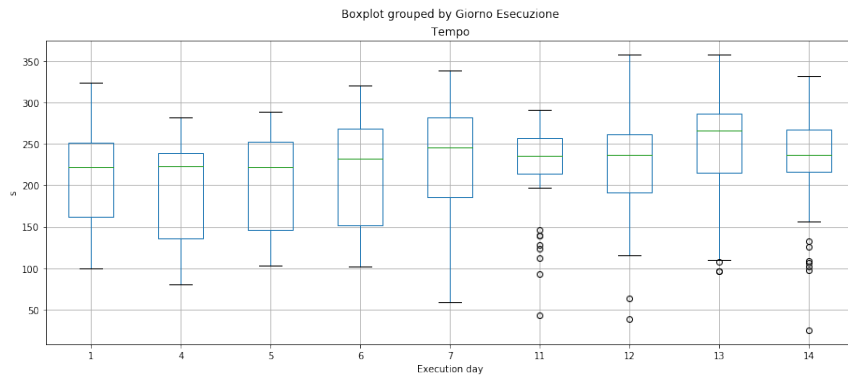
Since the visualization part shows that operation times are widely spread along the whole time cycle, the investigation goes towards the analysis of the order of operations made by the teams.

Since datasets are very poor, each median operation time, instead of mean operation time, is going to be compared among the operators in order to investigate if the operations sequence is respected or not, as already discussed above. Since the actual time targets are not available

3.3 Automatic Mura analysis



(a) Scatter plots for operation number 5 - Team A.



(b) Scatter plots for operation number 5 - Team B.

Figure 3.17: Scatter plots for operation number 5 for the two teams involved.

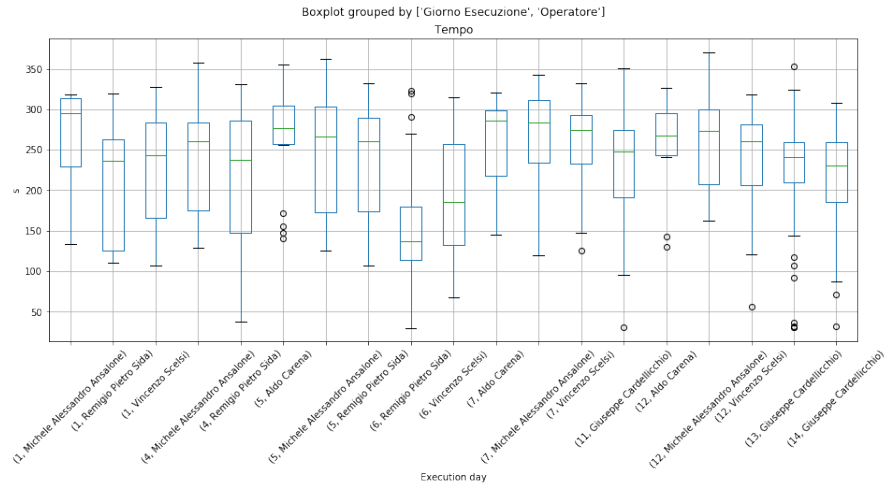


Figure 3.18: Caption

for security reasons, is not possible to evaluate directly time differences between ideal and actual operation times. Hence, a different approach has been used: if operations are performed within "timeThreshold" seconds each other, they can be considered not switched. The analysis is made for teams. If the real target operation times were available, it would be interesting to evaluate how many operators follow properly the operation tags and look for the process causes that are involved in operations switch. Listing 3.5 shows the criteria of decision of ordered/non-ordered operation.

With the used method, the 55% of operations are not ordered at all. This tool works whatever the threshold applied was because it is independent from the reference times. Whenever the target times were available, would be possible to consider it for the time difference.

Listing 3.5: Key method to evaluate whenever the two teams use to switch the operation execution with another one

```
def isOrdered(operation , timeThreshold):
    if (abs(operation["Team_A_median_time"] -
        operation["Team_B_median_time"]) > timeThreshold):
        return False
    else:
        return True

def orderedOperationsPerc(ops , timeThreshold):
    notOrdered = 0
    unorderedOperation = list()
    numOfOperations = len(ops)
    for operation in ops:
        # If the operation could be considered ordered add 1 to the counter
        if (isOrdered(operation , timeThreshold)):
            pass
        else:
            notOrdered = notOrdered + 1
            unorderedOperation.append(operation["OperationsName"])
    return 100*(notOrdered/numOfOperations), unorderedOperation
```

The previous analysis has been carried on in parallel with the WO pillar leader of Mirafiori

plant, and results were absolutely coherent. Basically they have every kind of information related to the assembly line. They used the whole set of MES relevant operations in order to effectively detect the switch of the order of operations. The actual scenario they found is the following (figure 3.19):

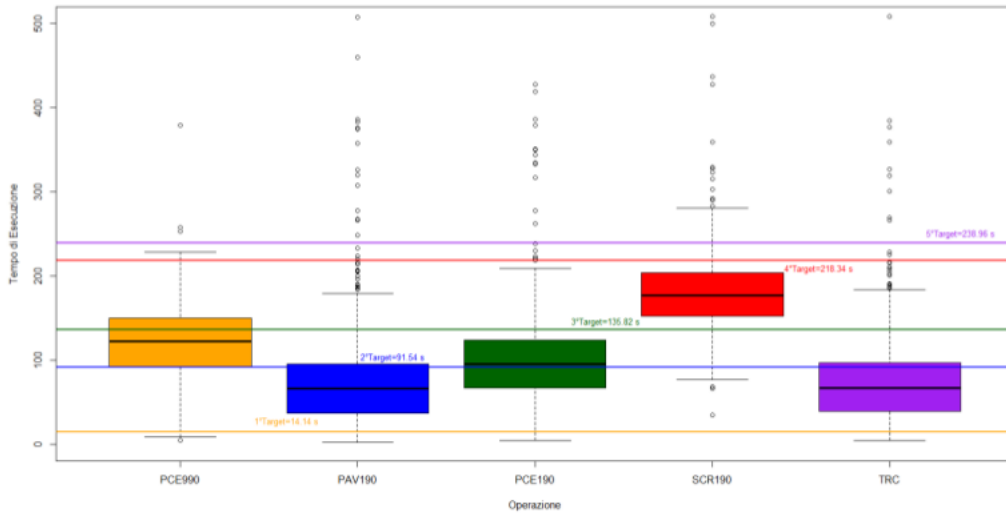


Figure 3.19: MES relevant operations for a single workplace. X axis the operations, Y axis the execution time.

Operations along x axis are to be done consequently, hence the y axis should grow up more or less linearly. Instead, the actual situation demonstrates that operators use to switch operations. In particular, this happens because workers should validate and objectify the operation, whenever it is done. Validation happens thanks to some wearable devices or some big screens where operator must select the operation already done.

This analysis and the one made by WO team, highlight that operators use to certificate actions, and then actually execute them. It represents a huge problem in process stability evaluation, because it may happen that a worker certifies an operation and at the end it is not completed at all, generating a line stoppage and a contamination in MES database.

Autocorrelation analysis

After the demonstration that some operations are completely random, a correlation coefficient has been evaluated in order to test how the two teams work linearly.

Since the most of operations are nor Gaussian, nor ordered in sequence coherently, another kind of activity to be performed is checking the correlation of the same operation made by the two teams.

It could be possible that operations are not ordered, but anyway operators work following the same logic and it might reflect on the fact that operations would have an high level of correlation. This analysis can be performed by using Pearson correlation coefficient that gives a measure of linear correlation between two variables X and Y (that represent, in this case, the distribution times):

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\rho_X \rho_Y}$$

where

- $\text{cov}(X, Y)$ is the covariance of the two variables, which is a coefficient of joint variability
- ρ_X is the standard deviation of X variable
- ρ_Y is the standard deviation of Y variable

Depending on Pearson correlation coefficient, the two variables may be correlated, not correlated or negatively correlated. The coefficient has a value between -1 and 1, where:

- $\rho_{XY} > 0$ variables are directly correlated
- $\rho_{XY} < 0$ variables are indirectly correlated
- $\rho_{XY} = 0$ variables are uncorrelated

From the official Python documentation: *The Pearson correlation coefficient measures the linear relationship between two datasets. Strictly speaking, Pearson's correlation requires that each dataset be normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases* [31]. The following method inserts Pearson correlation coefficient between the two teams operation times into the data structure, with the assumption that distributions are Gaussian. It was supposed that distributions are lognormal, hence in order to bring them to a more Gaussian behaviour, the method was applied to the logarithm of the time series.

The result is that every operation has a Pearson coefficient greater than 0.80, at least, meaning that the two teams work more or less coherently. However, is important to notice that the method may be not entirely reliable for datasets larger than 500 entries, for the reason already explained in the previous paragraph.

Data conversion

Original datasets present temporal features, but as they are, is not possible to use them because they are just strings. Hence, the data format has been introduced thanks to Pandas package and temporal features (columns) are added to the dataframe (figure 3.20)

Esecuzione	Tempo	Squadra	Tempo squadra A	Tempo squadra B	TempoLog	Ora Esecuzione	Turno	Giorno Esecuzione	Minuto esecuzione	GiornoSettimana Esecuzione
2017-12-01 06:09:53.017	227	A	227.0	NaN	5.424950	6	1.0	1	9	4
2017-12-01 06:15:58.227	181	A	181.0	NaN	5.198497	6	1.0	1	15	4
2017-12-01 06:24:49.443	298	A	298.0	NaN	5.697093	6	1.0	1	24	4
2017-12-01 06:29:00.563	188	A	188.0	NaN	5.236442	6	1.0	1	29	4
2017-12-01 06:34:16.137	169	A	169.0	NaN	5.129899	6	1.0	1	34	4
2017-12-01 06:40:19.917	201	A	201.0	NaN	5.303305	6	1.0	1	40	4
2017-12-01 06:46:06.830	217	A	217.0	NaN	5.379897	6	1.0	1	46	4

Figure 3.20: Example of dataframe with the updated features

Dataset consistency

Starting from this assumption, having as purpose to perform more in depth analysis, like applying machine learning algorithms, of making whatever kind of data aggregation, is necessary validating available data and test the usability of them. The instrument used to perform this task is Principal Component Analysis (PCA). Before performing any kind of machine learning analysis, is needed to evaluate the consistency of the current dataset features, i.e. be sure that information is not lead by a single feature, otherwise dataframe would not be useful. PCA provides also a method to reduce a complex dataset to lower dimension to reveal sometimes hidden, simplified structure that often underlies it. If a strong correlation between variables exists, the attempt to reduce the dimensionality makes sense. In a nutshell, this is what PCA is all about: finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.

In this case, PCA is used to check if every feature brings a certain amount of information with respect to the others by evaluating the related eigenvalues, which represent the percentage of variance of the selected components. The eigen vectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones can be dropped. If the whole amount of variance is brought by just a single feature, the dataset is not consistent and not usable.

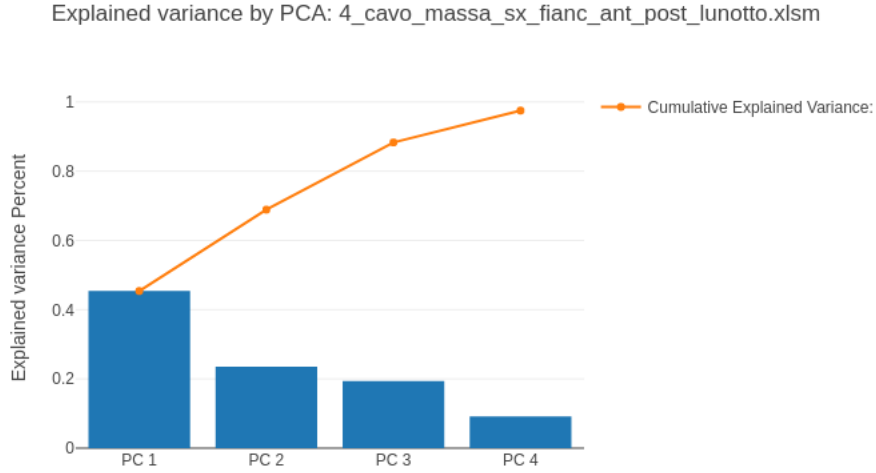


Figure 3.21: PCA applied on the dataframe of operation number 4

From a practical point of view, a useful measure of the how good is the dataset is the so-called "explained variance," which can be calculated from the eigenvalues. The explained variance tells us how much information (variance) can be attributed to each of the principal components. The following method shows graphically the cumulative explained variance against the percentage explained variance [43]. Figure 3.21 shows an example of PCA applied on the normalized dataframe of the operation number 4. Every operation analyzed has the same pattern, hence the datasets contains well spread information. Hence, it is no needed any dimensionality reduction or artificial data manipulation.

After being applied different statistical methodologies, found some useful insights in terms of operations switched, correlation between teams and visualized information that can be very useful to the people working in assembly line, like team leaders and supervisor, it is going to be useful apply some machine learning algorithms to find some hidden correlation and to exploit a feature of decision trees to select the most relevant column to the variability of the operation time.

3.3.2 Contributors detection

The datasets used have different columns, each of whose is related to the temporal features of the operation and to the operator itself. This paragraph focuses on finding a set of *contributors*

that have caused the variability in the operation time. The ideal situation would forecast that a mathematical model describes the operation: unfortunately the elements for building a robust model are not that large, that means that there is a not negligible possibility that machine learning algorithms fail the contributors detection. The output of this analysis is finding a set of "*relevant*" features that allow to build the model avoiding the use of meaningless features. This contributors detection phase is represented by a classification task and the consequent analysis of the model resulting from the testing phase.

Classification labeling

Every operation is translated into time window by some specialists in labour field and stored in a proprietary software called TiCon, as already explained in previous chapters. These operation times are considered as ideal references for each operation. The Mura is the variability of the executed operations with respect to the reference times. The following analysis focuses on investigating and clustering the main factors and reasons why the variability happens.

In the very early stage of the classification analysis, the problem is translated into a binary classification problem by separating the "in time operations" from "out of time operations" by selecting a percentage of the median time and subtracting from it to obtain the left side "out of time" region and adding to obtain the right side "in time" region. The median is used because it is more robust than the mean against the outliers and because of the dataset dimension. The result, assuming the distribution as Gaussian, is shown in the figure below (figure 3.6), where times distribution is splitted into three decision regions. The reference methods used are listed into 3.6.

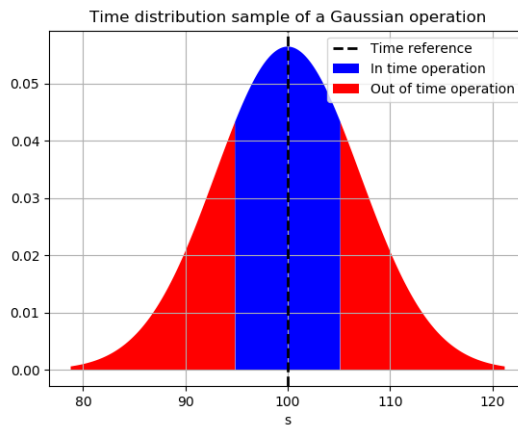


Figure 3.22: Qualitative example of classification for a (supposed) Gaussian distribution of operation times.

Where in particular, the labels are so defined:

- Out of time $\rightarrow 1$
- In time $\rightarrow 0$

Listing 3.6: Methods that perform classification based on the concept of "reference time"

```
def binaryClassification(x, meanTime, percentage):
    leftLimit = meanTime - int(meanTime * percentage)
    rightLimit = meanTime + int(meanTime * percentage)
    if ((x < leftLimit) | (x > rightLimit)):
        ris = 1
    else:
        ris = 0
    return ris

def insertLabels(dictionary, percentage):
    listaOperatori = dictionary["Lista_operatori"]
    for op in listaOperatori:
        dic = dictionary[op]
        meanTime = dic["Median_operation_time"]
        leftLimit = meanTime - int(meanTime * percentage)
        rightLimit = meanTime + int(meanTime * percentage)
        dic["Left_limit"] = leftLimit
        dic["Right_limit"] = rightLimit
        dic["Operator_dataframe_LGBM"]["Classification"] = \
            dic["Operator_dataframe_LGBM"] \
            ["Tempo"].apply(binaryClassification, args=(meanTime, percentage))

def insertLabelsPerOperations(dictionary, percentage):
    meanTime = dictionary["Dataframe"]["Tempo"].sort_values().median()
    #meanTime = dictionary["Median operation time"]
    leftLimit = meanTime - int(meanTime * percentage)
    rightLimit = meanTime + int(meanTime * percentage)
    dictionary["Dataframe"]["Classification"] = dictionary["Dataframe"]["Tempo"] \
```

```
.apply(binaryClassification , args=(meanTime, percentage))
```

In order to evaluate some classification parameters, the following method computes the classification's output versus the real classification labels: 3.7

Listing 3.7: This method computes the performance of the classification

```
def getClassicatorParameters(x, x_hat):
    "x= true classification "
    "x_hat= predicted classification "
    "1→ Out of time "
    "0→ In time "
    falsePositive = 0
    falseNegative = 0
    truePositive = 0
    trueNegative = 0

    for i in range(len(x)):
        if (x[i] == 1 and x_hat[i] == 1):
            truePositive = truePositive + 1
        if (x[i] == 0 and x_hat[i] == 0):
            trueNegative = trueNegative + 1
        if (x[i] == 1 and x_hat[i] == 0):
            falseNegative = falseNegative + 1
        if (x[i] == 0 and x_hat[i] == 1):
            falsePositive = falsePositive + 1
    accuracy = round((truePositive + trueNegative) / i, 2)
    precision = round(truePositive / (truePositive + falsePositive), 2)
    recall = round(truePositive / (truePositive + falseNegative), 2)

    return accuracy, precision, recall
```

The method returns basically three variables. They are the most used parameters in computing the performance of a classification task:

- **Accuracy** is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy

then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, consulting other parameters is needed to evaluate the performance of the model.

- **Precision** is the ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall** or sensitivity, is the ratio of correctly predicted positive observations to the all observations in actual class.

Another source of evaluation for a classification task is the so-called Receiver Operating Curve (ROC) [12]. It is a graphical schema for binary classifiers. In x axis, there is the false positive rate, that is number of false positives over all the entries, and in y axis there is the true positive rate, also called recall that is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both are defined between 0 and 1. In other words, this curve highlights the ratio between the correct detection and the false alarm, and it is prepared by setting a threshold value based on which choosing the prediction value. It is ideal to maximize the true positive rate while minimizing the false positive rate. The ideal situation would happen when the number of false positive is zero, and the number of true positive is one. Hence, the ideal position of the ROC would have (0, 1) as coordinates.

When the ROC is built, a parameter called Area Under the Curve (AUC) that evaluates the area under the ROC. The closer to 1 is, the better the classifier is. The AUC however, appears to be one of the best ways to evaluate a classifier's performance on a data set when a "single number" evaluation is required or an operational point has not yet been determined [5].

In the end, the last parameter used to discuss algorithms' performances is confusion matrix. It is a table that gathers some useful information in terms of true labels versus predicted labels. It is typically used for binary classification. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).

Machine learning algorithms

Basically, the classification task has been carried on using decision trees because of an interesting feature. Decision trees have a particular field which indicates how each feature is relevant with respect the others in the current classification task which is called *feature importance*. Basically, each leaf of the tree splits the dataset into two parts trying to maximize the information gain in among every possible split. Of course, the more a feature is used into a tree to split the data,

the more is important and relevant. Furthermore, decision trees require a little data preparation: they do not require data normalization because it deals with models that need absolute values for branching.

In the end, a prediction has been performed in order to check the model goodness. Since, again, the dataset is very small, different upsampling applications have been performed because is not assumed that datasets fits well with the decision trees. Hence, classification and parameters have been evaluated in different scenarios:

1. Upsampling of the whole dataset and using the same train and test dataset
2. Upsampling of the whole dataset and split train and test datasets
3. Upsampling only train dataset

Upsampling method There are a number of methods available to oversample a dataset used in a typical classification problem: the most common technique is known as SMOTE: Synthetic Minority Over-sampling Technique (SMOTE). Basically, each row belonging to the minority class has k neighbors in feature space; then, the new synthetic data is chosen among the neighbors multiplied by a random number between 0 and 1. From literature, is shown that applying this technique improves classification performances [4]. The upsampling procedure has been performed by using a Python package called *imblearn*.

The following paragraph deals with the discussion of the algorithm used.

LightGBM LightGBM is a gradient boosting framework that uses tree based learning algorithm, made within a project by Microsoft⁸. Gradient boosting is a machine learning technique of regression and classification that produce weak predictive models, typically decision trees. It builds the model through the optimization of a loss function with a certain criterion. Hence, basically boosting is a gradient descent technique that has the aim of building a classifier by gathering different weaker classifiers trying to parameterize them with some weight coefficients trying to improve the performance in gradient descent procedure [20].

For sake of clarity, to find weak learner, base ML concepts are applied with different distributions. Each time the learning rule is applied, it generates an output, which basically is a new weak prediction rule. This is an iterative process: after a lot of iterations, the boosting algorithm mixes

⁸<https://github.com/microsoft/dmtk>

the weak learners into a single stronger prediction rule, combining and assigning to each of them some weights that will be used by the final learner.

LightGBM is therefore a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm. Since it is tree based, it splits the dataset through its leaves, which are the weak learners, rather than splitting the data depth wise. Hence, when growing on the same leaf, in LightGBM, the leaf-wise algorithm can reduce more loss than the depth-wise one, and result is much better in terms of accuracy (Figure 3.23 shows the different concepts of level and leaf wise algorithms).

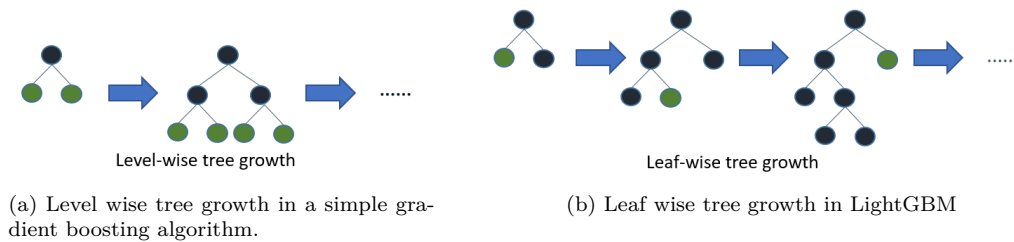


Figure 3.23: Diagrammatic representation by the producers of LightGBM to explain the difference between level and leaf wise algorithms.

Leaf wise leads to increase in complexity and may lead to overfitting: it may be overcome by defining some parameters related to the trees, like the max depth or the maximum number of learners (leaves).

Very recent literature shows the strength of this algorithm and the huge number of advantages it brings:

- Fast training and high efficiency
- Low memory usage
- Better accuracy than other boosting algorithms
- Compatibility with large datasets
- Parallel learning supported
- Reduces calculation cost of split gain thanks to the use of histogram to split data: replaces continuous values with discrete bins
- Supports categorical features with no need for encoding them

Unfortunately, it does not work that well with small datasets, like those used in this work: the resulting learners are, as discussed later, very overfitted to the data, hence a prediction task would result very complex, in this case.

In this work, the implementation of the algorithm is provided by Microsoft itself as an open source library.

Random Forest Random Forest is a machine learning algorithm that can perform whatever kind of tasks, but is typically used for regression and classification. It is based on bagging (Bootstrap Aggregating): is an ensemble algorithm designed to improve the stability and accuracy of machine learning methods in order to improve the diversity of learning models, introducing randomness in input data at the expense of the precision of the single learners.

Bagging is considered as a special case of model averaging approach, trying to introduce randomness in the choice of samples and learners in order to remove bias and overfitting. In fact, Random Forest, like LightGBM, has different learners (trees). In order to randomize the choice of classifiers leaf by leaf, the choice of the best feature on which performing data splitting is not made on the whole dataset⁹, but on an undersample of features.

In particular, Python package used¹⁰, performs an average on all learners

A *forest* allows to build a classifier made by lot of decision trees and that presents a result generated by single trees' prediction as output. The main feature is that each tree, basically works on an undersample of instances randomly extracted from the starting dataset. The more correlated are the trees, the higher will be the error. Hence, the purpose is to select independent features such that final classifier will have learners with very low correlation.

As in LightGBM case, Random Forest works well with large dataset. At least, with a high number of features. In this case, dataset is very limited, hence you would expect a very poor result in terms of classification.

Analysis report for cluster of operators The analysis described above, using the algorithms described and the upsampling technique performed on cluster of operators does not bring any relevant information. The classifiers' features do not lead to any useful insight: the most operations

⁹Bagging is important because without this one, every node would choose with a high probability the same variables to split data.

¹⁰Python package used for Random Forest is scikit-learn

are affected by the minute of execution, which is obviously not relevant at all. A possible way to use these data could be by producing a descriptive model able to describe the operations from a statistical point of view, but is very difficult to perform some kind of prediction based on the available data. For this reason, an additional analysis has been performed by clustering available data on operations, even if it is already been discussed that the variability introduced by different operators is very consistent. Hence, the expected result is that one of the main contributors would be operators, with the possibility to perform some predictions in terms of the most critical operators and the most critical stations.

For operations clusters, the same logic pipeline is applied in order to detect Mura contributors. In this case, analysis should be a little bit more complicated due to the categorical features "Operator" and "Team". While LightGBM deals with this kind of feature, for Random Forest algorithm an encoding process is needed. At the end, the logic flow of the analysis remains the same as well. In order to repeat the same analysis, it is needed to say that SMOTE, in Python does not support categorical data. Hence, an encoding procedure is needed for both LightGBM and Random Forest, even though the first one supports categorical data.

3.3.3 Upsampling of the whole dataset and using the same train and test dataset

The first step aims to check that algorithms fit into datasets features. If they fail this phase, then going on with the analysis would be totally useless. What is expected is very high values of AUC, almost close to 1. Whenever the classification is done, there is always an unbalancing in minority class. Hence, SMOTE technique rises the number of minor class entries and balance again the dataset improving the classification performances.

Figures 3.24 show the classification performances of operation number 2 when LightGBM is applied. Since both training and testing phase are performed on the same dataset, the performances are excellent in terms of sensitivity and specificity. More or less, the same result is got for the other operations.

A comment should be done for feature importance plot. Since the dataframes are very small and poor in features number, the result shown in 3.25c does not make that sense: the first feature in the plot is the one that is more present into the final model, but it is the feature with the highest level of variance, which can indicate that model highly overfits data.

The Random Forest algorithm outputs the same results. Hence, these algorithms may work properly with this kind of datasets.

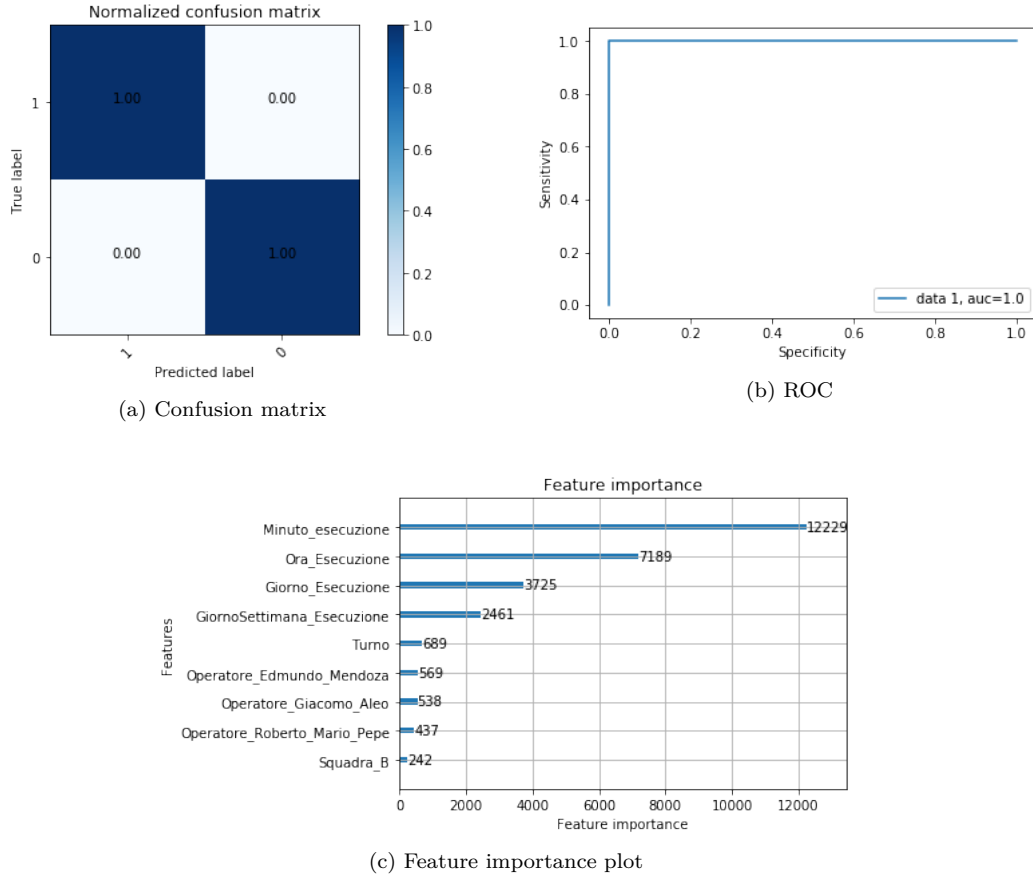


Figure 3.24: Operation 2: LightGBM performance analysis for training and testing on the same dataset.

3.3.4 Upsampling of the whole dataset and split train and test datasets

In this second phase, the dataset is totally upsampled using SMOTE, but training and testing steps are made on different dataframes. Figure 3.25 shows LightGBM results.

Again, the Random Forest algorithm shows almost exactly the same results. The table below 3.1 summarizes the values of accuracy, precision, recall and AUC of every operation per either LightGBM and Random Forest. Anyway, the feature importance graphs show that the most relevant feature is the one which has the highest variance. This fact, again, reflects the huge level of overfitting of models on the data, due to the limited number of entries.

Classification results seem now to be more realistic. Even if there is still the same problem of overfitting because of the datasets dimension. If more data would be available, a more robust testing procedure would be carried on and models would be validated. Until now, is just possible to describe operations through these overfitted decision trees.

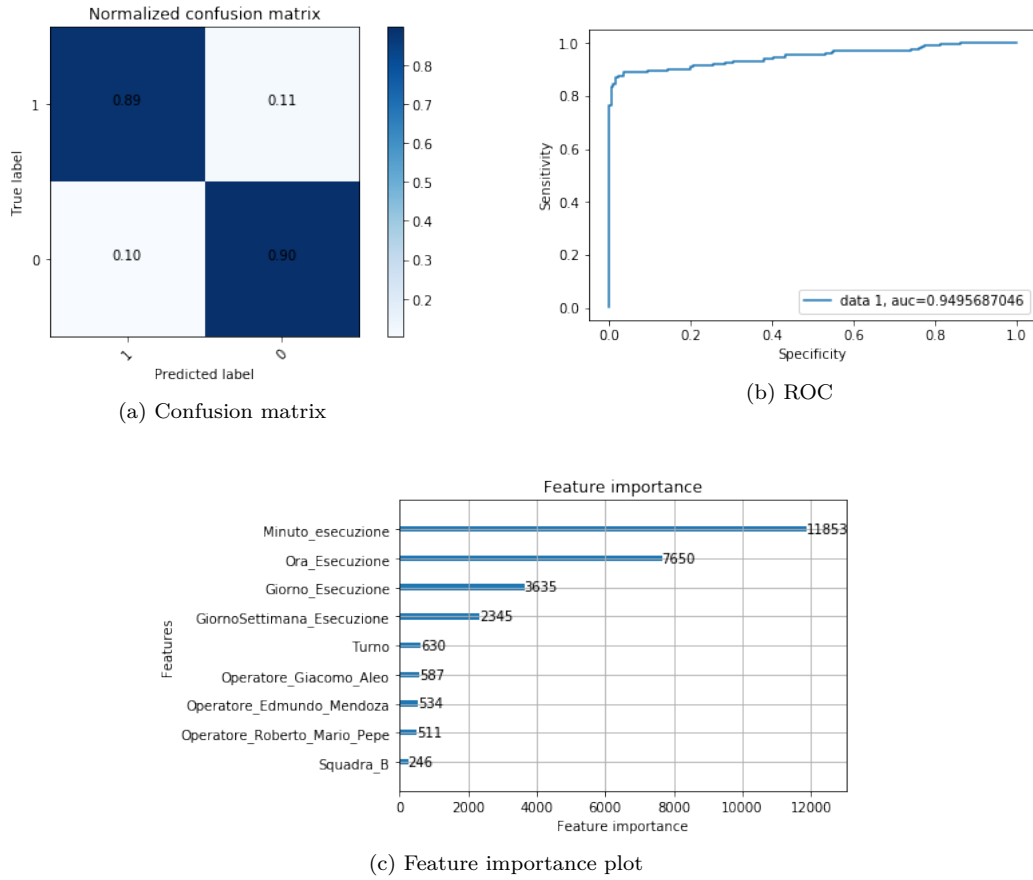


Figure 3.25: Operation 2: LightGBM performance analysis for upsampled dataframe. Training and testing phase made by splitting dataset.

Table 3.1: The table below summarizes the classification performances for the upsampled dataset.

Operation	Acc_LG	Acc_RF	Prec_LG	Prec_RF	Rec_LG	Rec_RF	AUC_LG	AUC_RF
Operation 2	0.89	0.92	0.88	0.94	0.9	0.88	0.94	0.94
Operation 6	0.79	0.79	0.75	0.75	0.85	0.83	0.84	0.84
Operation 5	0.78	0.82	0.71	0.78	0.86	0.83	0.87	0.89
Operation 23	0.64	0.64	0.62	0.66	0.63	0.5	0.68	0.63
Operation 1	0.93	0.92	0.93	0.95	0.92	0.89	0.96	0.96
Operation 4	0.71	0.69	0.72	0.71	0.64	0.60	0.74	0.74
Operation 3	0.8	0.79	0.8	0.81	0.79	0.75	0.89	0.86
Operation 5_2	0.65	0.65	0.66	0.67	0.64	0.62	0.73	0.71
Operation 4_2	0.86	0.92	0.85	0.92	0.85	0.85	0.95	0.97

3.3.5 Upsampling only train dataset

The last task is the one closer to the machine learning itself: the upsampling procedure is only applied to the training set in order to model the trees and then apply them to an actual testing dataset. The following table (3.2) shows how the performance of classifiers are drastically diminished in terms of AUC, especially.

Table 3.2: The table below summarizes the classification performances for the upsampled training set.

Operation	Acc_LG	Acc_RF	Prec_LG	Prec_RF	Rec_LG	Rec_RF	AUC_LG	AUC_RF
Operation 2	0.82	0.84	0.18	0.08	0.11	0.03	0.53	0.46
Operation 6	0.70	0.68	0.79	0.78	0.82	0.80	0.57	0.51
Operation 5	0.73	0.77	0.79	0.88	0.72	0.70	0.75	0.81
Operation 23	0.57	0.56	0.53	0.53	0.53	0.28	0.62	0.54
Operation 1	0.89	0.91	0.25	0.43	0.12	0.12	0.59	0.66
Operation 4	0.64	0.68	0.54	0.60	0.53	0.05	0.69	0.70
Operation 3	0.72	0.66	0.78	0.77	0.86	0.78	0.60	0.57
Operation 5_2	0.58	0.60	0.35	0.37	0.34	0.32	0.55	0.50
Operation 4_2	0.76	0.80	0.19	0.17	0.43	0.29	0.70	0.75

Partial results From the classification parameters table 3.2, is possible to understand that, at the moment, with the available data there is not the possibility of doing prediction on variability operation times. A lot of operations have an AUC between 50 and 60%, meaning that the prevision is a little bit better than random classification.

Separating the decision regions

There is a last chance to improve performances of classifiers, which is separating the decision regions, deleting the operation times that are within a certain threshold of the reference time, starting from the logarithm of the original operation times. The following graph shows an idea of removing the entries that are in the middle of the two decision regions (Figure 3.26).

Unfortunately, the performance trend follows the one analyzed in the previous chapter: if from one hand there is a better optimization of the decision regions, on the other hand the datasets are smaller in terms of dimensions. Therefore, the models are more overfitted on the data and are not usable for the purpose of contributors detection to the variability in most cases. Sometimes, for datasets bigger than 500 rows, the classes splitting improves the classification tasks for either LightGBM and Random Forest. The improvement is very small in terms of precision, accuracy and AUC, but it is undoubtedly a methodological starting point point for the next steps analysis.

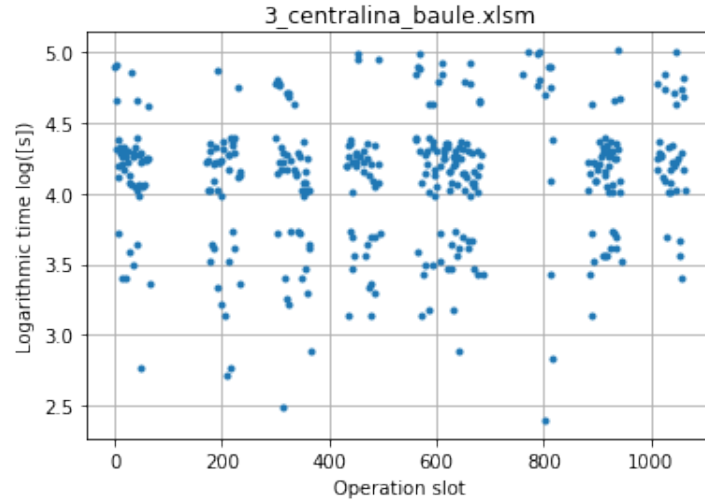


Figure 3.26: Scatter plot that shows the separation of decision regions. Time is logarithmic in order to allow a more balanced division.

Feedback to production line At the moment, this use case is subject of study of WO team of Mirafiori plant. They are going to perform analysis on MES relevant operation and they are going to use the visualization tools used in this work. What is not clear at all, is how to show the result of Mura analysis: initially, the aim was to build a real time algorithm able to detect Mura on operation times.

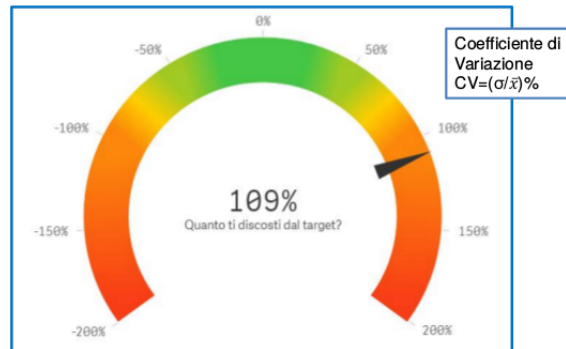


Figure 3.27: Caption

But it would be an example of over-engineering, because real time information would not improve that much line performances. Instead, it would be interesting to visualize the features of a shift detail, the most critical hours and the most critical operations. A possible indicator that summarizes variability is variance coefficient, which is a dispersion index that allows to compare different operation times without knowing the reference time. The figure 3.27 shows the indicator

that could be used into the screens in production line: it is a percentage coefficient and assumes that process is stable, therefore distribution times should be as much Gaussian as possible.

Chapter 4

Summary

In this chapter, there will be a detailed list of all results found thanks to the development of the use case about Mura analysis.

Analysis has been performed from two sides because of the reduced data size:

1. statistical analysis on true data
2. machine learning analysis on synthesized and manipulated data

Before analyzing results, is important to consider that during the internship in FCA, there have been a lot of difficulties in finding available data from the plants because of security reasons. This Master thesis, initially, was supposed to focus on big data and machine learning analytics and deal with IT systems in order to build a real tool which integrate with FCA systems.

Instead, people has proved to be very reluctant in sending operation data. Moreover, the teams working on Mura analysis, face huge bureaucracy processes to send data. Hence, a great objective reached by this work is the awareness of plant people and all stakeholders in new data analytics techniques potential.

Statistical analysis Statistical analysis has the aim of detect the major statistical features of the operations. At the moment, FCA plants use some parameters to evaluate and monitor the variability of operations. But, from literature review, the parameters that monitor process capabilities assume that events are normally distributed. Hence, the very first step of analysis aims to verify the distribution of operation times. The results emerged are that the 11.1% of available operations are Gaussian distributed. It is a very low percentage, considering the fact

that the analyzed operations are just nine, a small fraction of all MES relevant operations of the AGAP assembly line.

Once Gaussianity was tested, a big part of data visualization is performed to investigate the kind of distribution and looking for some parameters to keep under control and to use them as KPIs. Whenever the distribution features are investigated, there will be a graphical method to look for contributors to the variability.

Each dataframe is splitted into the two teams, because each team has its own operator and its way of working. Hence, an important result found is that each team works non-coherently with the other one. This suggests that there is a lack of rigour in working and in keeping the WCM standards. In fact, the 55% of analyzed datasets show the switch of the order of operations, without keeping the WCM standardization principles. The analysis would be improved if the whole MES relevant data of a station were available, in order to detect what operations are used to be switched. As it has been discussed in previous chapters, providing a tool able to detect the behaviour of operators, may help in breaking some wrong habits and improves the process methodologies.

Hence, the investigation goes more in detail into contributors detection.

The approach of visualization part is basically top-down, in the sense that original dataframe itself contains a lot of information and features. The criterion used is to unpack, step by step data in order to find interesting trends and find some correlations between time variability and operations' features.

Analysis starts with the plot of the frequencies of operation times with respect to the features, i.e. day of the week, operator, day of production, etc. Whenever interesting trends have been found, the final system would let the user to choose the column with which unpack data and zooming into data more in detail.

Basically, the main contributors found is operator. In fact, operators seem to work, sometimes, completely differently from the other one.

Machine learning analytics Machine learning analytics was aimed to face the problem of getting an automatic tool that outputs the list of contributors and provides predictions about variability.

In particular, this is a binary classification task, where classes deal with operation made "in time", and "out of time".

Basically, two algorithms have been tested on the datasets, divided per teams. The purpose of contributors detection has carried on by decision trees, chosen because of an interesting feature.

Algorithms chosen are LightGBM and Random Forest and they are applied, more or less, with the same mathematical parameters, e.g. number of nodes, depth, learning rate, etc. They by definition are composed by nodes, that are *weak learners*, because no matter what the distribution over the training data is will always do better than chance, when it tries to label the data. This means that the learner algorithm is always going to learn something, not always completely accurate, i.e. it is weak and poor. Each learner splits the dataset by a certain feature, and of course, the more the feature is present into the nodes, the more that feature is relevant to the classification.

Because of datasets size and the unbalancing of the classes, different setups are used in order to test how the algorithms work because a traditional approach seems to not lead any insights.

- Upsampling the whole dataset and applying algorithms to the same train and test set.
- Upsampling the whole dataset and applying algorithms splitting the resulting dataset into train and test dataset.
- Upsampling only the train set and apply the model to the test set as it is.

The first item works perfectly. Classification parameters are close to 1. This means that algorithms learns perfectly the trend of dataset.

The second one, for certain operations seems to work well (reference in table 3.2). This, again, depends strictly on dataset size. However, every model, suffers hugely of overfitting.

The last item deals with a more common machine learning pattern. For the reasons discussed above, performances are very bad: a lot of classifiers shows almost random prediction.

Since important results have not been found, the analysis went into deeper analysis of decision regions. In fact, from literature review, in classification tasks the more regions are separated, the better is the performance of prediction. So, another classification criterion has been applied to the dataset: the entries in an intermediate region, defined as a percentage of the median operation time, are not considered. Classification parameters improve for both the algorithms used. However, the mathematical model is not usable yet.

Finally, the actual output of this work is supposed to be very useful to the production line workers through the introduction, into the screens, of an indicator that updates shift by shift able to show variability in terms of percentage of displacement from the nominal value, per each operation.

Chapter 5

Future works

Since the context of this work, within the business environment, is at the early stage, there are a lot of fields to be improved into the future work steps.

Basically, the future works involves either the field of infrastructure system, and the setting of different data source integration and the application of machine learning algorithms in order to lead the manufacturing to more efficiency and higher performances.

First of all, with Industry 4.0 paradigms, it could be possible to integrate different IT systems in order to converge different sources into a single one. In this work, columns of dataframes seem to not bring relevant information. It would be interesting to integrate data from different fields like quality, external environment, market and business trend or other WCM indicators on people capabilities.

The integration between different data sources may be very difficult and complex, because each of them works in separate IT environments. So, the very first step towards the implementation of a robust architecture is building a cloud-like infrastructure, where every manufacturing datum converges and is properly stored. Then, an efficient system of API (Application Programming Interface) could be built in order to manage whatever kind of application, independently from infrastructure allowing scalability of the whole system.

Another aspect that has not been considered deals with customization of cars: each body, into the assembly line could have a lot of different customization accessories, This implies that operations may be slightly different from the "basic" operation. So, through machine learning analytics it would be possible to detect the most critical sequence of operations in order to better organize the flows of the bodies avoiding that more complex set of operations passes into a station. consecutively optimizing the cars' sequence.

A field in which WO teams are working on is the balancing of the line: in order to get an easier Mura analysis, lots of MES relevant operations are going to be shifted earlier in the time cycle in order to eliminate all the random and independent variability introduced by non-MES relevant operations.

In the end, some new KPI may be defined in order to keep under control the assembly line and operators: real time algorithms may introduce some kind of real time control and dynamic control on human made operations.

Finally, whenever the analysis is more robust is possible to implement an interface to the final users that get the tool usable. Actually, a mockup of the final interface is under construction.

Bibliography

- [1] Muhammad Aslam, Muhammad Mohsin, and Chi-Hyuck Jun. “A new t-chart using process capability index”. In: *Communications in Statistics - Simulation and Computation* 46.7 (2017), pp. 5141–5150. DOI: 10.1080/03610918.2016.1146759. eprint: <https://doi.org/10.1080/03610918.2016.1146759>. URL: <https://doi.org/10.1080/03610918.2016.1146759>.
- [2] WCM Association. *The zen of doing it better, and making it better*. URL: <https://world-class-manufacturing.com/Kaizen/kaizen.html>.
- [3] Jaiprakash Bhamu and Kuldeep Singh Sangwan. “Lean manufacturing: literature review and research issues”. In: *International Journal of Operations & Production Management* 34.7 (2014), pp. 876–940. DOI: 10.1108/IJOPM-08-2012-0315. eprint: <https://doi.org/10.1108/IJOPM-08-2012-0315>. URL: <https://doi.org/10.1108/IJOPM-08-2012-0315>.
- [4] Kevin W. Bowyer et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *CoRR* abs/1106.1813 (2011). arXiv: 1106.1813. URL: <http://arxiv.org/abs/1106.1813>.
- [5] Andrew P. Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. In: *Pattern Recognition* 30.7 (1997), pp. 1145–1159. ISSN: 0031-3203. DOI: [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2). URL: <http://www.sciencedirect.com/science/article/pii/S0031320396001422>.
- [6] Erik Brynjolfsson. *Data in Action: Data-Driven Decision Making in U.S. Manufacturing*. Jan. 2016. URL: <https://www2.census.gov/ces/wp/2016/CES-WP-16-06.pdf>.
- [7] G. J. Cheng et al. “Industry 4.0 Development and Application of Intelligent Manufacturing”. In: *2016 International Conference on Information System and Artificial Intelligence (ISAI)*. June 2016, pp. 407–410. DOI: 10.1109/ISAI.2016.0092.

- [8] Forbes - Louis Columbus. *10 Ways Machine Learning Is Revolutionizing Manufacturing*. June 2016. URL: <https://www.forbes.com/sites/louiscolumnbus/2016/06/26/10-ways-machine-learning-is-revolutionizing-manufacturing/#9635c3328c2c>.
- [9] O. L. DAVIES. *Statistical methods in research and production*. Oliver and Boyd, London, 1947.
- [10] Robert Dawson. *How Significant Is A Boxplot Outlier?* Feb. 2011. URL: <https://www2.amstat.org/publications/jse/v19n2/dawson.pdf>.
- [11] P. F. Dubois. "Guest Editor's Introduction: Python: Batteries Included". In: *Computing in Science Engineering* 9.3 (May 2007), pp. 7–9. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.51.
- [12] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: <https://doi.org/10.1016/j.patrec.2005.10.010>. URL: <http://www.sciencedirect.com/science/article/pii/S016786550500303X>.
- [13] Francesco Canuto - FCA. *World Class Manufacturing Evolution and Innovation*. Oct. 2017. URL: <http://www.convegnoaiig.it/2017/wp-content/files/2017/11/Francesco-Canuto-FCA.pdf>.
- [14] Fabio De Felice, Antonella Petrillo, and Stanislao Monfreda. "Improving Operations Performance with World Class Manufacturing Technique: A Case in Automotive Industry". In: *Operations Management*. Ed. by Massimiliano M. Schiraldi. Rijeka: InTech, 2013. Chap. 01. DOI: 10.5772/54450. URL: <http://dx.doi.org/10.5772/54450>.
- [15] Zahediasl S. Ghasemi A. *Normality Tests for Statistical Analysis: A Guide for Non-Statisticians*. Oct. 2012. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/>.
- [16] Google. *Tensorflow: An open-source machine learning framework for everyone*. URL: <https://www.tensorflow.org/>.
- [17] FCA Group. *Melfi plant: a high quality production line*. URL: https://www.fcagroup.com/plants/en-us/melfi/the_plant/Pages/default.aspx.
- [18] FCA Group. *WCM Development Center*. URL: https://www.wcm.fcagroup.com/en-us/development_center/pages/default.aspx#.
- [19] FCA Group. *WCM Development Center*. URL: https://www.wcm.fcagroup.com/it-it/wcm_at_fca/Pages/wcm_association.aspx.

- [20] Guolin Ke et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems* 30. Dec. 2017. URL: <https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/>.
- [21] Andrew Kusiak. “Smart manufacturing”. In: *International Journal of Production Research* 56.1-2 (2018), pp. 508–517. DOI: 10.1080/00207543.2017.1351644. URL: <https://doi.org/10.1080/00207543.2017.1351644>.
- [22] H. L. Monostori and E. Westkampfer. “Machine learning approaches to manufacturing”. English. In: *CIRP Annals - Manufacturing Technology* 45.2 (1996). ISSN: 0007-8506.
- [23] Todd D. Little and William W.S. Wei. *Time Series Analysis*. URL: <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898-e-022>.
- [24] Melissa Mallon. “Data Visualization”. In: *Public Services Quarterly* 11.3 (2015), pp. 183–192. DOI: 10.1080/15228959.2015.1060147. eprint: <https://doi.org/10.1080/15228959.2015.1060147>. URL: <https://doi.org/10.1080/15228959.2015.1060147>.
- [25] Politecnico di Milano. *PoliMi: Smart Manufacturing, la via italiana alla quarta rivoluzione industriale*. July 2016. URL: <https://www.digital4.biz/executive/polimi-smart-manufacturing-la-via-italiana-alla-quarta-rivoluzione-industriale/>.
- [26] Robert H. Mitchell. *Mini-Paper: Process Capability Indices*. URL: <http://asq.org/statistics/2010/09/quality-tools/process-capability-indices.pdf>.
- [27] Alena Otto et al. “Ergonomic workplace design in the fast pick area”. In: *OR Spectrum* 39.4 (Oct. 2017), pp. 945–975.
- [28] Pandas. *Python Data Analysis Library*. URL: <https://pandas.pydata.org/index.html>.
- [29] Fabian Schlötzer. *Industry 4.0 - The World of Smart Factories*. Sept. 2015. URL: http://studenttheses.cbs.dk/bitstream/handle/10417/5754/Fabian_Schlotzer.pdf?sequence=1.
- [30] scikit-learn. *scikit-learn Machine Learning in Python*. URL: <http://scikit-learn.org/stable/>.
- [31] Scipy.org. *Pearson coefficient - Python documentation*. URL: <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pearsonr.html>.

- [32] Nist Sematech. *What is Process Capability?* URL: <https://www.itl.nist.gov/div898/handbook/pmc/section1/pmc16.htm>.
- [33] Scot H Simpson. *Creating a Data Analysis Plan: What to Consider When Choosing Statistics for a Study*. Canadian Journal of Hospital Pharmacy, 2015.
- [34] Steven E. Somerville and Douglas C. Montgomery. “PROCESS CAPABILITY INDICES AND NON-NORMAL DISTRIBUTIONS”. In: *Quality Engineering* 9.2 (1996), pp. 305–316. DOI: 10.1080/08982119608919047. eprint: <https://doi.org/10.1080/08982119608919047>. URL: <https://doi.org/10.1080/08982119608919047>.
- [35] VDE Testing and Certification Institute. *German Standardization Roadmap*. Jan. 2016. URL: <https://www.din.de/blob/65354/f5252239daa596d8c4d1f24b40e4486d/roadmap-i4-0-e-data.pdf>.
- [36] Denis R. Towill. “Industrial engineering the Toyota Production System”. In: *Journal of Management History* 16.3 (2010), pp. 327–345. DOI: 10.1108/17511341011051234. eprint: <https://doi.org/10.1108/17511341011051234>. URL: <https://doi.org/10.1108/17511341011051234>.
- [37] Shiyong Wang et al. “Implementing Smart Factory of Industrie 4.0: An Outlook”. In: *International Journal of Distributed Sensor Networks* 12.1 (2016), p. 3159805. DOI: 10.1155/2016/3159805. eprint: <https://doi.org/10.1155/2016/3159805>. URL: <https://doi.org/10.1155/2016/3159805>.
- [38] Watson et al. *Detecting outliers in time series*. June 1996. URL: http://eprints.whiterose.ac.uk/2209/1/ITS261_WP362_uploadable.pdf.
- [39] Wikipedia. *Industry 4.0*. URL: https://en.wikipedia.org/wiki/Industry_4.0.
- [40] Wikipedia. *Process capability index*. URL: https://en.wikipedia.org/wiki/Process_capability_index.
- [41] Wikipedia. *The visual workplace*. URL: https://en.wikipedia.org/wiki/The_Visual_Workplace.
- [42] Thorsten Wuest et al. “Machine learning in manufacturing: advantages, challenges, and applications”. In: *Production & Manufacturing Research* 4.1 (2016), pp. 23–45. DOI: 10.1080/21693277.2016.1192517. eprint: <https://doi.org/10.1080/21693277.2016.1192517>. URL: <https://doi.org/10.1080/21693277.2016.1192517>.

- [43] Kazuyoshi Yata and Makoto Aoshima. “Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix”. In: *Journal of Multivariate Analysis* 101.9 (2010), pp. 2060–2077. ISSN: 0047-259X. DOI: <https://doi.org/10.1016/j.jmva.2010.04.006>. URL: <http://www.sciencedirect.com/science/article/pii/S0047259X10000904>.

List of Acronyms

AGAP	Avvocato Giovanni Agnelli Plant
AI	Artificial Intelligence
AM	Autonomous Maintenance
AO	Additive Outliers
API	Application Programming Interface
ARPANET	Advanced Research Projects Agency NETwork
AUC	Area Under the Curve
CD	Cost Deployment
CPS	Cyber-Physical System
DDDM	Data-Driven Decision Making
FCA	Fiat Chrysler Automobiles
ICT	Information and Communication Technology
IDE	Integrated Development Environment
IIoT	industrial Internet of Things
IO	Innovative Outliers
IoT	Internet of Things
IQR	Inter-Quartile Range

JES	Job Element Sheet
KAI	Kay Action Indicator
KET	Key Enabling Technologies
KPI	Key Performance Indicator
LSL	Lower Specification Limit
ML	Machine Learning
NaN	Not-A-Number
NoP	Number of Operators
NVA	Non-Value Added
OEE	Overall Equipment Effectiveness
OPL	One Point Lesson
PCA	Principal Component Analysis
PCI	Process Capability Indices
ROC	Receiver Operating Curve
SMOTE	Synthetic Minority Over-sampling Technique
SVA	Semi-Value Added
TC	Time Cycle
TIT	Total Industrial Technology
TPS	Toyota Production System
TT	Takt Time

USL	Upper Specification Limit
UTE	Unità Tecnologica Elementare
VA	Value Added
WCM	World Class Manufacturing
WO	Workplace Organization