POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Analysis of joint maintenance and quality control strategy in semiconductor fabrication

TESI DI LAUREA MAGISTRALE IN

MANAGEMENT ENGINEERING

Author: **Alessandro Baratelli**

Student ID: 953440
Advisor: Tullio Antonio Maria Tolio
Co-advisor: Maria Chiara Magnanini
Academic Year: 2021-22

POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Abstract

Nowadays the semiconductor manufacturing system is claiming for more predictive engineering tools to give a prognostic view of the manufacturing system equipment's health. In fact, degradation of a component/ system is one of the major factors that cause defective product output. Unfortunately, the complex process dynamics characterized by this sector doesn't allow to predict some unobservable degradation tools. This is primarily due to a number of pressing issues, including the fragmented data sharing between inspection, maintenance, and operation control, the limited and unreliable phenomenon of semiconductor processes, the weak or nonexistent correlation between equipment information and quality results, and the general lack of historical data. This is the reason why in semiconductor industry is still needed traditional solution to screen out output quality independently from preventive equipment control conditions. An innovative way is to combine these two approaches in order to integrate these two management practices (Advanced process control (APC) and SQC) for finding the optimal policy to exploit the overall information of the field to minimize the resources employed. . This thesis maintains an emphasis on this comprehensive framework to comprehend how the two control policies interact exploiting different information feedback given from the machine controller and the quality inspection results. How the decision-maker would rank these two control visibilities according to the production system design chosen on the shop floor is the key point of interest of the study. This comparative analysis is built around a Discrete event simulation (DES) and Response Surface methodologies models that analyze the zero defect manufacturing performances of a degrading lithography process in series with the overlay metrology station

In order to comprehend how the control policy interacts in various settings. This study demonstrates that there isn't a single solution for every production system configuration for the two control policy decision-making processes, because different production conditions lead to different information feedback accuracy from machine control and remote inspection.

Key-words: Yield rate, Throughput effective, Throughput total, Preventive maintenance, predictive maintenance, condition based maintenance, Advanced process control

Sommario

Al giorno d'oggi il sistema di produzione di semiconduttori rivendica strumenti di ingegneria predittiva per dare una visione prognostica dello stato di salute delle apparecchiature del sistema di produzione. Infatti, la degradazione di un componente/ sistema è uno dei principali fattori che causano la produzione di prodotti difettosi. Sfortunatamente, le complesse dinamiche di processo caratterizzate da questo settore non permettono di prevedere alcuni strumenti di degradazione non osservabili. Ciò è dovuto principalmente a una serie di questioni urgenti, tra cui la frammentazione della condivisione dei dati tra ispezione, manutenzione e controllo del funzionamento, il fenomeno limitato e inaffidabile dei processi a semiconduttori, la correlazione debole o inesistente tra informazioni sulle apparecchiature e risultati di qualità e la mancanza generale di dati storici. Questo è il motivo per cui nell'industria dei semiconduttori è ancora necessaria una soluzione tradizionale per schermare la qualità dell'output indipendentemente dalle condizioni di controllo preventivo delle apparecchiature. Un modo innovativo è quello di combinare questi due approcci al fine di integrare queste due pratiche di gestione (Advanced process control (APC) e SQC) per trovare la politica ottimale per sfruttare le informazioni complessive del campo per ridurre al minimo le risorse impiegate. . Questa tesi mantiene un'enfasi su questo quadro completo per comprendere come le due politiche di controllo interagiscono sfruttando il feedback di informazioni diverse fornite dal controller della macchina e dai risultati dell'ispezione di qualità. Il punto chiave di interesse dello studio è il modo in cui il decisore classificherebbe queste due visibilità di controllo in base al design del sistema di produzione scelto in officina. Questa analisi comparativa si basa su modelli di simulazione di eventi discreti (DES) e di metodologie di superficie di risposta che analizzano le prestazioni di produzione di difetti pari a zero di un processo di litografia degradante in serie con la stazione di metrologia di sovrapposizione.

Al fine di comprendere come la politica di controllo interagisce in varie impostazioni. Questo studio dimostra che non esiste un'unica soluzione per ogni configurazione del sistema di produzione per la decisione delle due politiche di controllo-processi di fabbricazione, perché le diverse condizioni di produzione portano ad una diversa precisione di feedback delle informazioni dal controllo della macchina e dall'ispezione remota.

Parole chiave: tasso di rendimento, tasso di produzione effettiva, tasso di produzione totale, manutenzione preventiva, manutenzione predittiva, controllo di processo avanzato

1 Sommario

1.1. The industrial context

The manufacturing sector is one of the most important key factors for nation's growth. Nowadays, companies are struggling in a highly competitive scenario, with mutating, fast moving and customer driven market. In the last decade some important global megatrends are affecting the future corporate strategies and more broadly the current living standards of modern societies.

In the last 20 years, the economic structure and system have led to an increase in the level of consumption of natural resources. The emerging developing countries have also contributed to higher living standards for most of the global population. Additionally, even the population growth rates will decrease in future years, at the end of the 21st century the population will be roughly around 11 billion [1] . This scenario will entail higher production capacity to achieve and higher resource and energy consumption for keeping the same living standard for an even larger population. For this reason, a decoupling of the consumption of material and energy from the rising global demand is required[2], [3]. Otherwise, the high consumption level will for sure lead to the reduced availability of virgin material and the generation of more and new wastes.

An example of the problems that this might create in Europe, could be the increased scarcity of raw materials used in High-Tech applications and consequently a price increase. This could threaten Europe Energy transition towards renewables. The decrease in material availability would increase the dependency of Europe on resource-rich countries like China or other countries worldwide [4].

International institutions and governments in Europe have translated this threat into an opportunity to develop new business models with scientific circular approaches together with incentives aimed to change the way of conducting business towards more sustainable business models. The Sustainable Development Goals (SDGs) outlined in the United Nations Global Agenda for Sustainable Development in 2015 [5] aims to the achievement of 17 goals within 2030. Therefore, the need for sustainable manufacturing has become stronger than ever and tools able to find and correct inefficiencies in the production system are fundamental for achieving this goal.

For this reason, manufacturing companies are facing the challenge of delivering the required production rates of high-quality products while minimizing the use of resources. The increasing emphasis on sustainable production requires maintaining the resource efficiency and effectiveness along the product, process, and production system life cycle.

The deployment of increasingly complicated designs to enable the administration and control of production processes has improved recently thanks to digitalization. Particularly, the Manufacturing Execution System (MES) has become important to business operations as the primary software module for the implementation of cutting-edge Zero-Defect Manufacturing techniques (ZDM). The implementations of MES have enhanced the control of manufacturing systems for production qualityNew sensor technologies in particular have made it possible to collect a wide range of data in real time on manufacturing lines while the process is being carried out at a very fast collection rate. To build a manufacturing system that is adaptable and customer-focused, the issue will be proactive control with appropriate data collection and integrating solutions. These variables result in a transition away from inflexible mass production toward an agile process that can respond with a minimal amount of changeover and production gap cost while always meeting the volume and quality requirements. To strike a balance between efficiency and effectiveness, the production system's complexity must be improved from a global perspective and at various levels. Manufacturing businesses are researching Quality, Production Planning, and Maintenance as essential processes that must be monitored in manufacturing systems to prevent suboptimal improvement in order to meet these objectives.

1.2. Motivations

Semiconductor Manufacturing system is one of the most complex manufacturing processes that consists of four basic steps: wafer fabrication, wafer probe, assembly (packaging), and final testing[6]-[7], [8]. The wafer manufacturing phase, often known as the front-end, is the most costly. During this phase, circuits are stacked onto the wafer using sequential procedures. There are numerous processing steps involved in this. The dynamics, performances, and characteristics of the process and the end output are thus determined by an unlimited number of factors. Since this market is subjected by fast-changing conditions, some structural reconfigurations, actions of improvement, or operational modifications must be taken into consideration analysing all possible alternative comparisons to design the optimal system for all the possible scenarios.

Generally, a number of analytical techniques have been developed to characterize a manufacturing system's behavior using equations that can aid in making precise decisions during production planning strategies[7],[9],[10],[11]. However, Analytical models are complex engineering formulations not easy to derive and not always they mirror the real behaviour of the system, since some restrictive hypothesis might not be aligned with more complex dynamics and improvement programs. Therefore, the introduction of Simulation tools plays an important role for the study of more articulated problems as it provides a closer interval of outcomes with respect to the real performances, even if it might be time consuming both for the design of the simulation of the real system and the extrapolation of the result.

Nowadays, the increasing competitiveness of the global market has resulted in a constantly increasing pressure on both the quality of products and the productivity of the systems. Therefore, the performance of semiconductor fabs is constantly being evaluated. Machine utilization, work in progress (WIP), flow time (FT), factory throughput, on-time delivery, and overall equipment efficacy are just a few examples of the conventional metrics employed. A portion of the research also looked at how they interacted and what impact they had on one another. A company's philosophy and the product market determine the importance of factory performance measures. Among all the performances that need to be considered, the quality control system represents a relevant factor for the design of the manufacturing plant. As described in the previous section, Manufacturing Companies wants to cope continuously the required production rates of good quality product with minimum waste of resources.

In the semiconductor industry, Yield is a crucial measure of manufacturing performance and equipment condition can have a significant impact on it. This performance is considered as the mean portion of die on a wafer that can ultimately be sold [6] .

Research and practice have usually dealt with the scheduling of equipment maintenance and production dispatching issues separately during the last few decades, neglecting the potential effects that equipment condition may have on various product categories or families. The issue is based on a scenario that occurs during the creation of semiconductor wafers, in which the equipment's quality degrades with time and negatively impacts the production process's yield.[12].

Scientists and entrepreneurs have never considered an integrated framework that combines Quality, Process Control, Production Planning, and Maintenance[12]. Instead, they have always studied these four areas independently. Though it's possible that improving one component at the expense of another will result in a less-than-optimal outcome. As a result, the success of a corporation that is based on production quality is heavily influenced by these three control policies, which work in concert.

Several empirical studies have been discussed about further interactions among the three-control policy. In particular, from a survey approach[13], potential correlations between the application of Just in Time (JIT) and Total Quality Management (TQM) lean practices in automotive and electronic industries are studied. These corporations have been able to better manage their production via higher-quality performance and reduce inventory using JIT strategies, producing beneficial results for their industry. The importance of comprehending the cause-and-effect relationships of the primary quality, production logistical, and maintenance factors has been underlined by this positive association. In [12] a Casual Loop Diagrams (CLD) has been delineated to consider manufacturing and shop floor related aspects. This aggregated representation model aims to highlight bi-directional mutual cause-effect relations found among quality, maintenance and production logistics and identify many existing trade-offs. The proposed CLD model can be seen as reference framework to describe the results of the work proposed in the following sections and to find unexplored problems that contributes for the improvement of effective throughput.

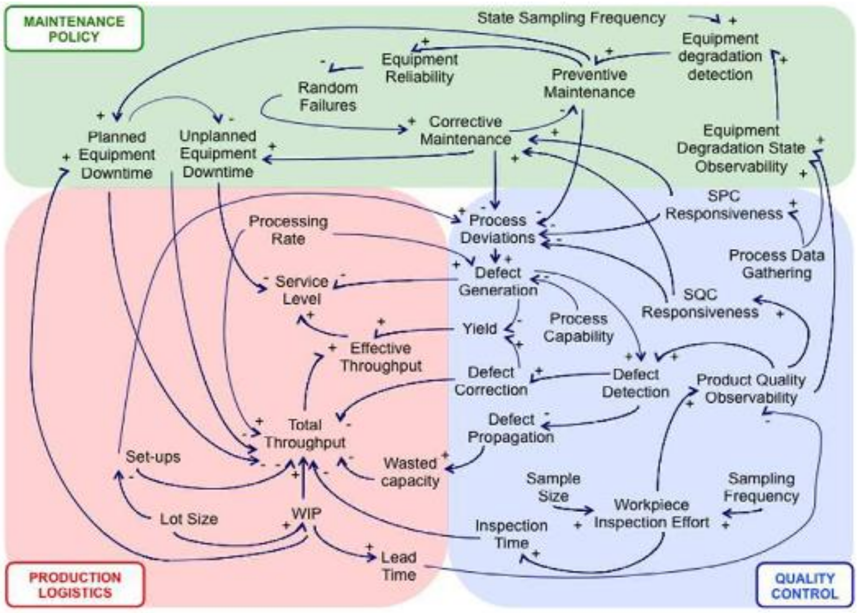


Figure 0.1: Details of Interaction model Casual loop diagram

As shown in the figure above, typical common dynamics are already known about manufacturing systems. A few significant interaction trade-offs are shown in the diagram: Inventory is reduced, which allows the system to immediately detect any quality issues. It is difficult to monitor processed components that pass through many stages and buffers before being inspected, and input for quality control is delayed. The buffers' ability to decouple the various machine speeds and avoid operational disruptions, however, has a favorable impact on the system's output rate. Low WIP may also hinder maintenance efforts since other stations may have an impact on the machine. Additionally, it has been demonstrated that the operating speed is negatively connected to the product's quality.[14] Thus, improving the machine processing rate has a positive impact on the system throughput, but may negatively affect the system yield.

Other interactions may be connected to the quality control architecture and its effects on productivity performance. These interactions must be assessed from the perspective of process control as well. The positioning of the measurement stations for the SoV-based robust control system and the timing of any corrective action made on the system from distant signals may have a considerable impact on the production flow limitations[9].The frequency with which resources are maintained improves part quality but reduces the operational time of machines, which has an impact on overall production. Although more inspections might be logical to better assess the degradation state of resources, this will increase the time it takes for the system to reach completion.

These problems usually involve different companies and different departments within each company. The coordination and cooperation among them in achieving a right balance between these conflicting goals is seen as a key issue for success. In the literature[12], The majority of works only include contributions pertaining to the integration of production logistics and quality interaction. Little progress has been made toward the three control policies' combined design. Fewer submissions from prestigious international publications address issues within a fully integrated framework, with the majority focusing on the interplay between quality and production logistics. These factors drive more research initiatives and creative suggestions for the behavior of production systems as well as for lowering costs for businesses.

A particular unexplored area in the semiconductor industry concerns the deep relation of equipment state condition and the quality result of the product. Manufacturing sectors such as aeronautical, railway and automotive, machinery, have already exploited machine learning tools, analytical and empirical studies to predict and infer respectively the future health condition of the equipment and the quality characteristic determined by the predicted level of degradation[15],[16],[17],[18].These tools are empowered by an optimal and cost-effective allocation of sensors to get real time information about the equipment system behaviour to proactively make maintenance intervention at the minimum expense, obtaining the best quality outcome. However, the high complexity of the semiconductor fabrication, characterized by hidden dynamics within different in-chamber chemical production stages [19]-[20], entails high efforts for the implementation of predictive and preventive maintenance policies (PdM , PM), because high investment of sensing tools is needed and the extraction of reliable data is still difficult[21]-[20]. Run-to-failure and periodic maintenance scheduling are still used in this area to determine whether systems are maintained well or poorly. Moreover, inspection stations like overlay metrology are considered the critical point for the major performance of the semiconductor line, since they usually deserve of the longest processing time to guarantee the certified quality level and lot of capital costs are implemented in order to not lose efficiency. In this stage SPC analysis and several types of intra-field and extra-field errors are examined to study whether the process is in control, the parts examined are not defective and if a correction of upstream processes must be taken in order to match some process deviations[7]. Indeed, the quality control plays an important role also as condition-based maintenance tool (CBM) for possible corrective actions along the upstream line. Nevertheless, this analysis misbehaves whenever there are long WIP, long buffer capacity and long serial process before the inspection for the reason already explained above. If the processed parts have to cross several production stages and buffers before being inspected a delay in the quality control feedback is generated and its responsiveness is reduced. These considerations strongly motivate the need for research activity and give an idea of its potential impact in terms of knowledge of production system behaviour as well as in terms of cost reduction for companies.

1.3. Objective

The Objective of this thesis is devoted to the analysis of possible interactions between the maintenance policy enacted by estimated equipment degradation from partial sensing of its real condition and the information quality related to the metrology control for an asynchronous two-stage manufacturing system.

The work fits in the framework described in 1.2 and tries answer to the following question: “Are both control policy useful to support synergically the joint consideration of quality, production logistics and resource maintenance” and “ When does the quality claims assumes predominant importance than the health estimation of the equipment?”

To make this analysis a simulation model is designed to describe a serial line composed by unreliable machine subjected to unobservable deterioration that influence with an unknown relation the product quality characteristics. An in situ controller that approximately guesses the behavior of equipment health condition using a hidden Markov model (HMM) is considered to handle a hypothetical forecasting tools approach of deterioration states of the machine. The simulation tries to emulate the semiconductor manufacturing system case proposed in [7], in which simulation is developed to validate the Approximate Analytical model as a support for the study of robust multistage process control model for measurement point reduction in the overlay metrology. A line's bottleneck is the inspection station, which must align many printed layers with subnanometric precision.

The new simulation model formulated will consider the dynamics of the case study to determine an appropriate maintenance policy that can be integrated with the SPC quality control that monitors the quality deviation from the normal average.

Furthermore, an optimization model through response surface methodology will be framed to conceive a production, quality and maintenance control policies able to minimize the long-term total cost, satisfying the production quality demanded by the final customer.

The model formulated will be a tool to understand how the system, will react towards changing conditions like the different knowledge of degradation states from the monitoring sensor, the different buffer capacity, different machine speeds and sampling rate.

1.4. Thesis outline

The thesis is structured in the following chapter:

- In chapter 2. A literature review of the main tools and topics treated in the thesis will be given.
- In chapter 3. A briefly recap of the Response surface methodology
- In chapter 4. The problem addressed by the thesis is formalized
- In chapter 5. Some problematic tied to the process degradation in the semiconductor process fabrication is presented to justify the simulation model construction
- In chapter 6. The simulation model to build is descripted
- In chapter 7. The simulation functioning is presented in detail
- In chapter 8. The discussion of the final result are formulated
- In chapter 9. Conclusion about the analysis conducted are drawn

2 Chapter two

2. Literature review

2.1. Notion and Definition of performance measures

Examples of concepts and initiatives that weren't designed for such rapidly shifting situations include Six Sigma, Just in Time, Continuous Improvement, Total Quality Management, Toyota Production System, and World Class Manufacturing. To achieve this goal, a modern, integrated conception of manufacturing quality must be created. There are several distinct Key Performance Indicators (KPIs) that each relate to quality, manufacturing logistics, and maintenance in the literature and in various business applications.[22]

The production rate, WIP, flowtime, defect rate, and other complicated non-linear functions of a single process or stage are common in manufacturing systems. The total quality management as a result (TQM) and total productive maintenance (TPM) paradigms are used in agile businesses and provide integrated KPIs to assess the success of a particular improvement plan's execution in an industrial setting. It has been demonstrated that there is a strong and favourable link between TQM and TPM that enables the implementation of practices simultaneously enhancing production performances.[23]-[22]. Particularly, the effective throughput—the total number of conforming components that the system produces over time—is shared by the two techniques as the most important integrated performance. The TQM philosophy, which pushes businesses to create products and set up operations that only produce commodities that match the expectations of the client, is built around this performance. This would guarantee that their resources are used effectively and efficiently to generate just the goods and services that the clientele want and is prepared to pay a premium price for.

Yield is another significant performance metric in the semiconductor business, and the health of the equipment plays a big part in determining it. According to [24] is the mean portion of die(chips) on a wafer that can ultimately be sold. Usually, three sites on the plant are used to calculate the measurement: The yield is first determined at the moment where the wafer fabrication process is complete by comparing the number of wafers that completed it to the number of wafers that began it (usually called Line-Yield). The second point is that, following wafer sort, the yield is determined by the proportion of tested dies that perform properly (usually called Die Yield). The third point is that, once the packaging and final testing procedures have been completed, the yield of that part may be determined by dividing the number of dies that pass the final tests by the number of dies that began the packing process. [6]

$$Overall\ yield = \frac{Q_{wafers\ out}}{Q_{wafers\ started}} \times \frac{Q_{good\ dies}}{Q_{dies\ on\ wafer}} \times \frac{Q_{dies\ pass\ final\ test}}{Q_{dies\ start\ packaging}} \qquad (1)$$

In section 4 our analysis will be done for the first contribution along the semiconductor fabrication.

Another concept of yield used by practitioners that is similar to the previous definition is the ration between the effective and the total throughput as described in the equation 4.21 in [9].

The problem is based on the situation found in semiconductor wafer fabrication where the equipment condition deteriorates over time, and this condition affects the yield of the production process.

One performance studied in literature associated with yield is the Flow Time (FT), which us defined as the elapsed time between the start and completion of a task [25] .

FT reduction is highly valued by semiconductor makers. By simplifying the system and creating control mechanisms for more effective line balancing, dispatching, and utilization, they hope to lower FT. With shorter FTs, a company may be able to fulfil client orders more quickly and respond to the market. Additionally, when FT declines, it becomes easier to spot process flaws, which speeds up the creation and improvement of the process. As a result, the company may increase its yield more quickly. In order to maintain their competitiveness, semiconductor makers therefore strictly regulate cycle times and work tirelessly to cut them. The procedure also shows the part's yield. [26].

Another important aspect that affects the quality control is related to the timing of the feedback about the quality condition monitored, that is considered in Lean manufacturing culture one of main cause of muda (wastes). Some common wait time is caused by processing delays, machine or system downtime, response time, or signature required for approval wait time.

If we break it down to its simplest form, a delayed response is a gap in time between when an event occurs and the response to that event. The gap in time is waste – waiting – and it is tied closely to the efficiency (or inefficiency) of your operations. [27]

All these performances are accounted in the work proposed, to deep analyse the behaviour of the system toward the new policy introduced.

2.2. Reference architecture for the industrial implementation of zero-defect Manufacturing strategies

A production system known as a Multi-stage Production System (MMS) consists of several parts, stations, or steps that must be finished for the final good or service to satisfy consumers' needs. A wide range of modern production and service systems are included in multi-stage systems, which are often employed in practice. Because MMS has a waterfall characteristic, the product's quality is influenced by both the results of earlier stages as well as the present stage. This presents great challenges for quality monitoring because of the large amount of data and the interactive effects of many factors on the quality of the product. Nowadays, recent technological developments provide the tools needed to understand and resolve these challenges.

Manufacturing firms have recently taken major steps toward digitizing software structures for managing and controlling production systems. Sensors, data collecting systems, and computer networks have become more accessible and affordable as a result of recent innovations, compelling industries to adopt high-tech approaches.

ERP, an outdated kind of production management software, is no longer sufficient to continuously monitor, regulate, and enhance the manufacturing system. The ability to efficiently deliver the required quantity with the required quality while keeping resource usage to a minimum level necessitates a thorough understanding of the operating manufacturing system, which can be attained by continuously gathering data and comprehending behavior to implement the most efficient control strategies.

In discrete manufacturing processes, total inspection at every intermediate step and extremely high sampling rates are already commonplace, and the volume of data gathered enables effective quality control system production execution. The large amount of Big Data has demonstrated that every stakeholder level must be taken into account for Multistage Manufacturing System (MMS) management to be effective.

Therefore, MES has acquired relevance as central software modules for the application of advanced Zero-Defect Manufacturing solutions, going beyond traditional data mining approaches.

In the following Figure, the proposed architecture within a manufacturing company in [28] is presented in detail .

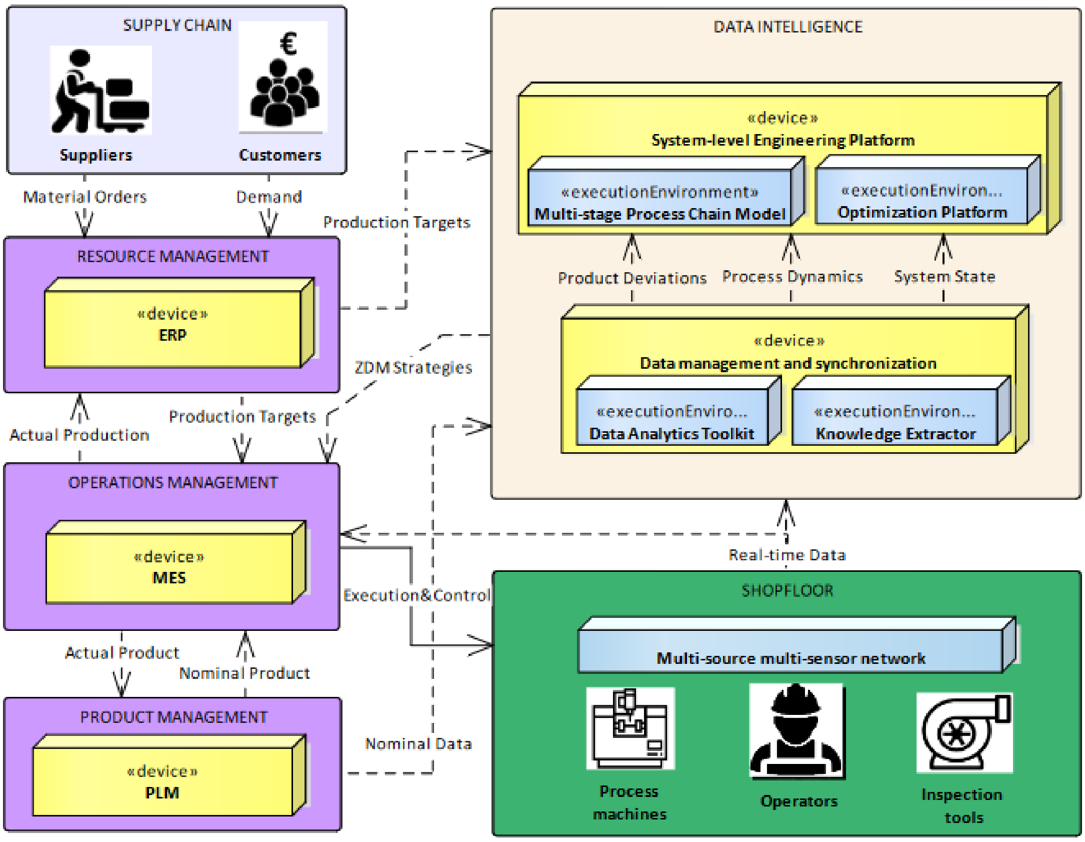


Figure 2.1: Reference architecture and software modules for ZDM strategies, with information flows (dashed arrows) and actuating controls (straight arrows)

The key management software that most businesses typically use is shown by the violet areas: Product Lifecycle Management (PLM), which contains basic product information such CAD files, a bill of materials (BOM), and production cycles (CAM); Manufacturing Execution System (MES), which actually manages operations at the shop-floor level, and Enterprise Resource Planning (ERP), which often manages information about incoming materials from suppliers and exiting goods to consumers. The systems of interconnected computational entities known as Cyber Physical Systems (CPS) that provide and use data-accessing and data-processing services made available on the Internet are what make it possible to use all of the aforementioned software. CPS are connected to the physical world and its ongoing processes.

Within the shop-floor area, the three main sources of data and information are represented by process machines, inspection machines and operators. These elements are markers for actual shop-floor machinery with field level sensors and actuators. Data Intelligence illustrates the suggested architecture, which is intended to relate to the existing architecture in order to develop and implement ZDM strategies.

The [28] identifies three layer necessary to build the overall architecture:

- The hardware layer composed by multi source and multi sensor network aimed to monitor machines state, inspection phase and operators.
- The data management and synchronization layer connect the hardware to the architecture's applications using appropriate wrappers such as OPCUA or XMPP protocols. Data Analytics tools and Extract Knowledge valuable for ERP, PLM, and MES devices are used to collect relevant data.
- The Engineering Platform also intends to provide a system engineering platform that can integrate maintenance, production logistic control regulations, and quality considerations into a unified framework. The assessment is based on information received by the preceding layer on process states, dynamics of material flow, and quality conditions. The ZDM paradigm's last phase is crucial since it calls for the development

of three different kinds of control loops: a low-level control loop that examines the current condition to avoid failure or errors. a middle level that assesses potential compensation or rework strategies if problems arise. Last but not least, the high degree of control that forecasts the viability of the quality/logistics solutions developed at medium level, as well as the economic advantages of those solutions.

To this reason, in the literature wide interest has been given to the analysis of production line performances from a productivity point of view and both analytical and simulation-based tools have been developed. However, less attention has been paid to the study of the relationships between quality performance measures, process control and productivity of the production systems. in [12] is highlighted the need for joint consideration of quality, production planning, and maintenance and proposes production quality as a new paradigm that goes beyond traditional six-sigma approaches.

2.3. Data management structure and architecture in the semiconductor fabrication

Data is the fundamental basis for all other Predictive and Preventive maintenance functions and applicationsManufacturing semiconductors has always required a lot of data. But in a "just in case" situation, the data is only transferred into storage. Additionally, more than 90% of the processed data is never again accessible. The following are the primary data sources used in this sort of plan: [21]:

- *Fault Detection and Classification (FDC)* The semiconductor industry has acknowledged data as a crucial element of advanced process control (APC). Status variable identification (SVID), which comprises equipment categorization with timestamp occurrence and problem detection, is collected during the semiconductor production process. With this knowledge, engineers may quickly return the machine to its initial state by referring to the machine fault number.

- *Equipment tracking system (ETS),* collects and analyses the equipment status and the proceeding operation states together with initial timestamp and fish timestamp described by operator.

-*Metrology data,* which records wafer measurement along the wafer-n in the lot-m.

Consolidation and synchronization of the three different datasets are required. In doing so, wafer-n and time stamps are used to combine events data, trace data, and metrology data. Anyway, the multidimensional information presented by this integrated data may be synthesized. To make the data more suitable for exploratory analysis, the dimensionality of the data has been reduced using a variety of techniques, including Principal component analyses (PCA). A class of such techniques often preserves the relationship between the data and projects it onto a low-dimensional space, either linearly or non-linearly. The Self organized Mapping(SOM) is a set of unique methods that reduce the amount of data by clustering and reduce the dimensionality of the data through a non-linear projection data on a small space [19]-[21]. Once this information are synthesized and cleaned, they are ready to feed the IT Architecture described in section 2.2. in figure below is shown a possible structure of semiconductor information system architecture

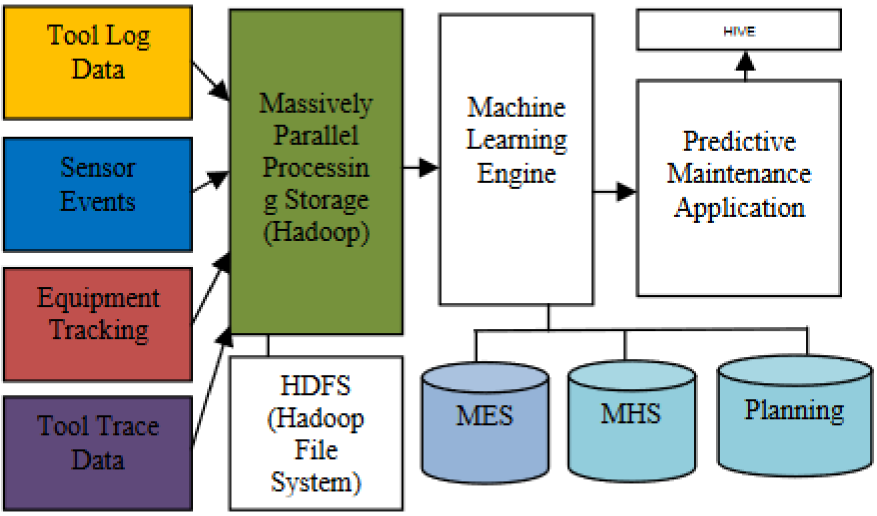

Tool Log Data
Sensor Events
Equipment Tracking
Tool Trace Data
Massively Parallel Processin g Storage (Hadoop)
HDFS (Hadoop File System)
Machine Learning Engine
HIVE
Predictive Maintenance Application
MES
MHS
Planning

Figure 2.2: conceptual big data architecture[21]

2.4. Inspection planning in a multistage system

In multi-stage system the design of an effective and cost-efficient quality control strategy pass through an optimal inspection policy. Inspection planning entails determining the part quality checks, as well as the location and scope of inspection activities in the production system, as well as the multi-sensor system to be installed for process monitoring. This are the main factors discussed in [29]- [12]to assess the significant quality characteristics of products while maximizing the system efficiency. In order to perform a machine and process state diagnosis and implement corrective or preventive actions to restore in-control manufacturing system behavior, the resulting decision will require the use of data gathering systems to provide useful information to SQC, SPC, and Condition Based Maintenance (CBM) procedures.

Generally, inspection can be done in every production stage, and the measured parts can be scrapped or even remanufactured according to the inspection result. [12] highlights how scrap activities can waste more energy, time, and materials, but that the lack of defect checking along the phases necessitates more rework and repair procedures to restore an item.

In literature, different analytical models as [7], [9] have studied the performance of the two different types of inspection stations: In a production line, machine can be monitored locally or remotely.

The figure Represent a local (M2) or remote (M1) inspection

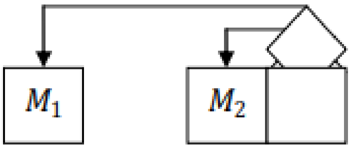


Figure 2.3: description of remote and locally inspection station

In the first scenario, if every machine is regulated locally, there is less chance that non-conforming products will get through the MMS, even though the cost of inspection may be higher than the savings realized by spotting flaws early on. Although there may be a non-conformance in the finished product corresponding to the first step, in the second scenario the cost of the inspection operations is decreased. This often happens because of Quality or failure correlation of the upstream process which may affect the quality outcome or even the integrity of the downstream process [12]. Moreover, [7] outlines how the remote solution generates further non-conforming items simply due to the delay of responsiveness from when an upstream stage out of control occurs till when it is detected at the end of the line. In fact, during this time interval pieces between the inspection station and the upstream stages can be judged as non-conforming item to be scrapped or reworked. To this reason, [9] has studied an optimal allocation of a limited number of inspection stations in the serial manufacturing system with the objective of maximizing system effective throughput. It illustrates that a higher level of effective throughput may be obtained with fewer inspection stations that are widely distributed rather than more stations that are badly distributed, meaning better performance for less money invested, particularly when the quality issue is serious.

The inspection can be performed over the entire lot or within a partial sampling. Researchers usually consider the inspection accuracy 100% perfect for sick of simplicity but not always in reality this may happen. Whenever an item is recognized as non-conforming product, four possible conditions can be applied. In fact, the item can be reworked, repaired, replaced or scrapped [29].

Manufacturers can save money and time by comparing the minimal number of items to be inspected to a 100% inspection. Sampling inspection also enables the evaluation of a large number of test items that would otherwise be impossible to inspect in a 100% inspection.

However, sampling inspections do not guarantee the quality of all manufactured products. Therefore, sampling inspections need a system that cuts inspection costs while taking into consideration the manufacturers and consumers' benefits, while reducing the risk of nonconformity to ensure consistently high quality. This means that the inspection scheme and how the appropriate sample size for the manufacturing process is determined are very important. Different joint optimization models such as [30] takes already into account an Average outgoing quality level(AOQL) that denotes in the optimization the number of defectives observed by the final customer.

The samples are often subject to the first type error and second type error, which are respectively mistaken rejection or missed rejection inside a test of hypothesis applied over time on the samples to verify whether the process is in control. These two errors applied on the samples are considered probabilistically by most of the researchers as an important tool to manage the uncertainties about inspection tools and inspection operators [31], [32].

Usually, to optimize an inspection planning, minimization of the total expected cost or the expected unit cost is used as objective function [29], however no study has considered in minimizing total manufacturing time and there is a general lack of multi-objectives models [33]. Researchers have devised a wide range of methodologies for resolving inspection planning difficulties, most notably with a non-linear total cost function optimized with gradient-based methods [32], but also with simulations, even if this method only considers a small number of production scenarios [33]. It is critical to design multi-objective optimization frameworks that include inspection and maintenance tasks simultaneously. To preserve profitability and worldwide competitiveness, manufacturing organizations must achieve a high degree of quality in their services or goods [31].

In fact, the production capacity is impacted by the maintenance operations required to return the used equipment to excellent condition. If production is run properly, it may result in high-quality products, and if the equipment is kept in good condition, it can run faultlessly. When checking the quality of parts, maintenance procedures should be considered. It makes appropriate to inspect less frequently as time goes on and/or the operating stage degrades when inspection costs are high[34] [32]. has presented a model for integrated planning of the part quality inspection and preventive maintenance activities while production stages are deteriorating. [30] has demonstrate a tight correlation between Condition Based Maintenance Policy or Preventive maintenance action with the various inspection strategy as inspection allocation and sampling inspection. [35] the quality control strategy proposed in this paper consist of a derivation of a continuous sampling plan proposed by that randomly inspects a fraction f(.) of products with 0<=f<=1[30]. However, in contrast of assuming a constant fraction of inspection, a dynamic sampling is assumed to incorporate the effects of a degrading process with continuous deterioration of parts quality [30], [36] The countermeasure proposed is the increase of sampling inspection size as the machine wears.

2.5. Work in progress scrap

Reworking defectives and managing waste or trash are significant difficulties in a manufacturing system that demand prompt attention in order to achieve the fundamental goals and conditions of a lean production system. A perfect lean system would be free of errors or defects at all levels, however this is unavoidably impossible. Therefore, scrap formation must be included when assessing the effectiveness of production systems because it is likely unavoidable. In reality, the parts produced by downstream machines until the issue is detected must be thrown away if a production line includes in-line inspection and the item is determined to be faulty. The cause may be an OOC machine[7]. Thus, the line has to be unloaded before the failed machine is repaired. Another case where material need to be scrapped along the line during production is when dealing with goods whose physical or chemical characteristics fall out of specification during stoppage. The food sector, for example, is a good example of where specific procedures must be completed in a timely and carefully controlled manner, and where extended failures and disruptions in the manufacturing process can result in significant quality degradation[37].However this phenomenon is not accounted in this work, the pieces trapped in the buffer are not unloaded after OOC of the process occurs.

Additionally, a major cause of in-line scrap is the prolonged exposure of material to certain conditions (such as heat, humidity, acidity, etc.). Numerous other production processes also include stoppages that might cause WIP to be damaged and ultimately need to be trashed. Neglecting the influence of scrap in assessing system performance might result in a significant approximation on the engineering side of manufacturing systems when long linear production lines are reviewed inside the system. In the literature [38] considers a two-workstation model where, in the event of a workstation malfunction, the part it contains is discarded. [39] The author examines a transfer line with geographically distributed waiting periods and downtimes, in which the component is discarded with certain probability if a workstation fails.[40] evaluates a line with two workstations and no intermediate buffer. One part can be stored at each workstation. When a workstation malfunctions, the malfunctioning component is eliminated as soon as the workstation is functional once more. In all of the previous research, it was presumed that if a workstation malfunctioned, the one component on the workstation would be thrown away. More recently [41] Create a model for a buffer-less, timed, automated transfer line where, when a workstation fails, it shuts down along with all the other workstations upstream, and the components stuck in the halted workstations are discarded after a predetermined length of time. In [42] The effect of quality on system productivity is investigated, but there is no scrap in the system, and the number of defective components is calculated at the end of the line. In [7], As soon as an OOC is detected, an online inspection machine stops the unit that produced the faulty item and utilizes Statistical Process Regulate (SPC) to control other processes. After that, without passing through the complete line where the yield is calculated, the material caught between the production and inspection equipment is scraped.

2.6. Maintenance policy classification

Maintenance strategy is a planned way to upkeep devices, which contains actions such as identification, researching and execution of many repairs, replace and inspect decisions[43].

Historically, several Authors and different country standards [43] has interpreted and classified maintenance techniques in different ways. [44] based on the strategies considered with conventional maintenance factors, a categorization was suggested. They present reactive and preventive maintenance as the two primary strategies in machinery maintenance based on this classification, as well as a variety of tactics related to maintenance concepts under the heading Proposal of Maintenance-types Classification to Clarify Maintenance Concepts in Production and Operations Management. According to Figure 2.4 and with the addition of Tables 2.1, 2.2, and 2.3. The authors draw the conclusion that it is clear that each of these strategies may be applied using various approaches, methodologies, and technology.

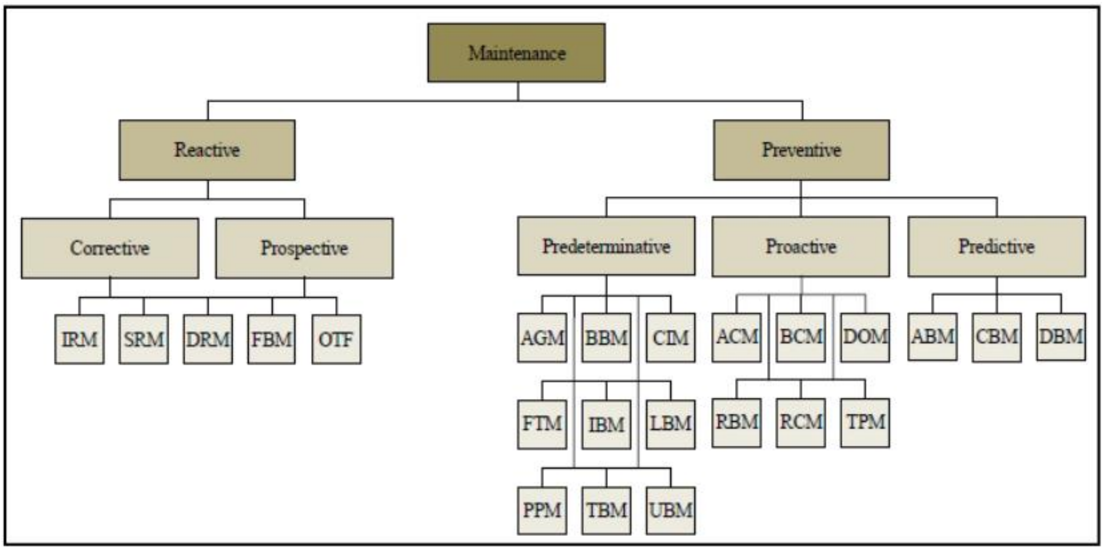


Figure 2.4: Maintenance taxonomy

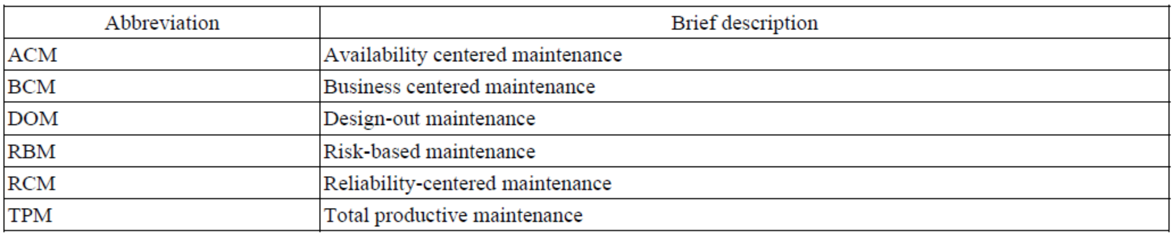
Table 2.1: Reactive tactics in Maintenance

Abbreviation	Brief description
IRM	Immediate reactive maintenance
SRM	Scheduled reactive maintenance
DRM	Deferred reactive maintenance
FBM	Failure-based maintenance
OTF	Operate to failure

Table 2.2: Preventive tactics in Maintenance.

Abbreviation	Brief description
AGM	Age-based maintenance
BBM	Block-based maintenance
CIM	Constant interval maintenance
FTM	Fixed time maintenance
IBM	Inspection-based maintenance
LBM	Life-based maintenance
PPM	Planned preventive maintenance
TBM	Time-based maintenance
UBM	Use-based maintenance

Table 2.3: Proactive tactics in Maintenance



Abbreviation	Brief description
ACM	Availability centered maintenance
BCM	Business centered maintenance
DOM	Design-out maintenance
RBM	Risk-based maintenance
RCM	Reliability-centered maintenance
TPM	Total productive maintenance

More in general, all these techniques are categorized in four maintenance techniques:

Instead of performing preventive maintenance, it was sometimes decided to run the equipment until it broke down. *Corrective maintenance*, also known as *reactive maintenance*, results in unscheduled downtime, higher labour costs, and faster asset obsolescence. On the one hand, the costs of routine maintenance are continuously declining, but in the long term, breakdowns or accidents cause protracted downtime and a fixed replacement cost. As a result, failure might happen suddenly or as a result of overusing the system. As a result, numerous pieces of equipment are updated often without being fully evaluated for usability or performance. All of these issues, together with technological advancements in the fields of computers, high-precision sensors, and low-cost sensors, drove the development of predictive maintenance forward (or condition-based maintenance).

Preventive maintenance ((PM) is a time-based statistical technique for anticipating and avoiding equipment breakdowns. In order to increase an asset's remaining useful life (RUL) or identify an asset that has reached the end of its useful life and is about to fail or break down, a set of tasks are performed at a set of intervals that are determined by the passage of time, the volume of production, and the condition of the machine. A PM policy has also been taken into account for a degraded system with a respectable level of dependability.

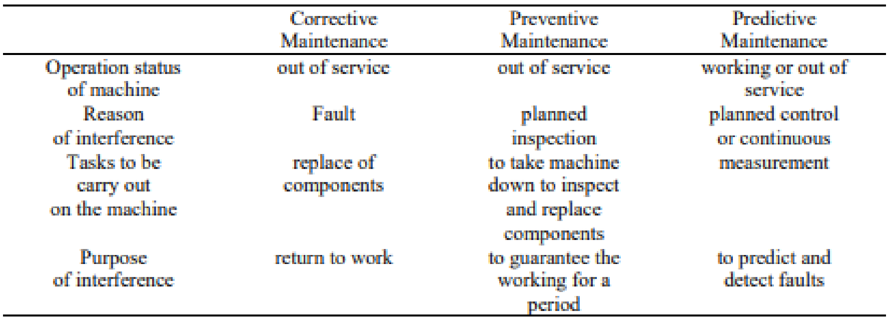
Despite all of its advantages, preventative maintenance comes with a number of difficulties, such as the requirement to create complex scheduling processes, extremely limited hours, as well as dangerous and expensive components. In addition, there is the issue of unanticipated failure outside of the schedule, which would render the machine or piece of equipment inoperable until the scheduled date. Employee safety is at stake, and replacing items that could have lasted longer and didn't need to be replaced could cost a lot of money. carrying out pricey emergency repairs or subsequent upkeep. Preventive maintenance has been the focus of all of these challenges and issues.

Predictive maintenance. According to a system's condition, predictive maintenance determines whether or not to maintain it. It is based on applying a number of non-destructive tests to characterize the machine's state and choose when maintenance is necessary.

Condition-based maintenance (CBM). The PM service is based on a reading or measurement that exceeds a set limit. If a machine is unable to maintain a tolerance, condition-based maintenance is performed.

Using a predictive maintenance program has several advantages. It can save downtime for the equipment, increase asset reliability, and prevent needless maintenance. Instead, it might result in a more complex system design by raising the fixed cost of purchasing diagnostic equipment, sensors, and software. A sound maintenance plan is an important and vital part of management in firms since it decreases failures, saves costs, and boosts productivity. It may be difficult for businesses to choose a workable strategy. As a result, many businesses employ a variety of maintenance strategies.

Table 2.4: Maintenance characteristics



	Corrective Maintenance	Preventive Maintenance	Predictive Maintenance
Operation status of machine	out of service	out of service	working or out of service
Reason of interference	Fault	planned inspection	planned control or continuous measurement
Tasks to be carry out on the machine	replace of components	to take machine down to inspect and replace components	
Purpose of interference	return to work	to guarantee the working for a period	to predict and detect faults

However, Maintenance activities are traditionally considered in conflict with production operations [35]. [44] Preventive maintenance protects equipment from degradation, decreasing the need for difficult and costly corrective measures, yet also has a detrimental impact on equipment availability. As a result, the attainment of production targets is threatened. To mitigate this unfavourable effect, substantial effort has been put into developing rules to improve the synchronization of preventive maintenance and production processes. Nowadays the Concept of Opportunistic maintenance was introduced in [45], but applications in industry have been very limited, mainly due to the difficulty in calculating the duration of opportunistic windows, in predicting the effect of the specific maintenance actions on the system. In general, the opportunistic maintenance is a maintenance intervention performed during a favourable opportunity time window in which the preventive maintenance on the machine won't be detrimental for the performance of the system.

In [46]Passive and Active Opportunity windows are defined:

- Passive windows exploit the idle time caused by the downtime of another machine in the system as happens in starvation and blocking phenomena
- Active windows exploit the inventory stored in the buffers to absorb a minor intervention on the machine, without interrupting the material flow in the system [47], [48]

Although the mentioned papers have contributed to the formalization of the problem, the definition of an optimal opportunistic maintenance policy is still missing.

2.7. Joint maintenance and quality strategies

Degradation can occasionally just mean an increased likelihood of failure. However, one of the main causes of faulty product production is component or system deterioration. As a result, implementing a preventive maintenance plan on the component or system to keep it in excellent condition and in compliance with the anticipated product quality criteria is a classic approach for lowering the number of faulty units. The output can also be sampled in order to check for damaged devices. It is revolutionary to combine these two methods in order to integrate these two management principles for determining the best course of action while reducing the total anticipated cost. The combined application of the combined application of SQC techniques and PM methods for achieving higher product quality and more effective use of resources has been investigated at single machine level [12], [49], [50] .Later [51], [52]combined the two approaches, at system level.

EWMA charts are often used as a SQC tool in the semiconductor industry, with the goal of serving as a typical Run-to-run (RtR) control in the semiconductor manufacturing process. In [53] a Double Exponentially weighted moving average (dEWMA) is simulated to demonstrate the performance benefits in the semiconductor system. [54] describe all possible applications od EWMA and MEWMA charts in semiconductor processes to signal multivariate deviations to the upstream processes. This tool can be integrated with policy maintenance proposing an optimal optimization rule.

In [55] an optimal adaptive control policy for machine maintenance and product quality control is derived. Moreover, [56] developed an optimal process control and maintenance procedure under general deterioration patterns, and [57] minimized the cost of an integrated systemic approach to process control and maintenance based on EWMA control charts by using genetic algorithm.

A performance measurement system for integrated SPC and CBM procedures is proposed in [122]. These works show that quality control based on product measurements can be useful for enhancing improved maintenance procedures. [58] developed a model for statistical quality control with an integrated optimization-based maintenance model for multicomponent series systems using an exponentially weighted moving average chart

In order to provide a complete and integrated model of quality and equipment failure propagation dynamics at the system level, these models may be used in correlated multi-stage production systems. Preventive maintenance would be used to increase quality robustness. Only the quality robustness of a system with out-of-control equipment has been addressed by simple machine reliability models (single state model). High levels of service are made possible through preventive maintenance, which also affects how well produced components are. When properly implemented, preventive maintenance techniques can lower production variation and thereby enhance service quality. As a result, models need to take into account both quality robustness and preventative maintenance.

2.8. Equipment deterioration Prognosis and Advanced control loop

High-tech manufacturing systems are getting more and more complex nowadays, and data collection and generation is happening far quicker than data processing. Continuous data collecting necessitates data analysis that is both new and efficient. With better Integrated Circuit (IC) design and production technology, the semiconductor industry has gathered a variety of data sources for mining operational knowledge. To gain this knowledge, the manufacturer must promptly and efficiently analyse the supplied data in order to operate the equipment at maximum efficiency and achieve a high production yield. Manufacturing faults such process variations and unanticipated tool failures are the main causes of poor tool utilization and low production yield. Unexpected failures have more catastrophic repercussions and significantly raise investment and operating costs in the semiconductor sector. The average cost of a 300 mm fab for producing 25,000 wafers per month exceeds US$ 2.0–3.0 billion [59]. Such colossal investment urgently necessitates the improvement of operational effectiveness. The cost of the equipment often makes up the largest portion of the total capital expenditure among all other costs. High tool utilization and top equipment effectiveness have therefore become essential objectives for semiconductor manufacturers with considerable capital investments in equipment. On the other side, the complex manufacturing environment has made it more difficult to adapt machine settings and has worsened production errors such process variations and unexpected tool failures. Corrective tactics such as "fix it when it breaks" were the oldest and most typical strategy for preventing such violations into in-situ control at semiconductor production[60].The problems with this strategy are various as the occurrence of unexpected breakdowns at inconvenient periods of production.

This phenomenon leads to uncommitted friction and wears in production schedules that directly results in market profit loss and customer dissatisfaction. Instead, CBM as a paradigm in Advanced Process Control (APC) theory infers equipment condition and alarms required action based on the runtime data analysis [61].

For this purpose, CBM requires a prognostic module that represents the healthy state of the equipment’s behaviour, which enables a manufacturer to avoid equipment breakdown and unnecessary maintenance, and a diagnostic module to identify the causes of the equipment failures.

These two prognostic and diagnostic modules aid in the development of sophisticated equipment degradation modeling and monitoring, which not only produces effective equipment condition monitoring but also aids in the identification of any failure causes. Minimizing scrap wafers, lowering unscheduled equipment malfunctions, minimizing unqualified periods of the equipment, and thus maintaining high process yields require such a model to be developed in the quickest time possible. In a nutshell, the fundamental motives for prognostic and diagnostic equipment deterioration modeling and monitoring are to reduce repair and maintenance expenses as well as associated operational disruptions, and to maximize tool use and production yield. Figure 2.5 shows the application of the prognostic and diagnostic results in the industry that takes place well in the current

Advanced Process Control (APC) system in semiconductor manufacturing proposed by [59].

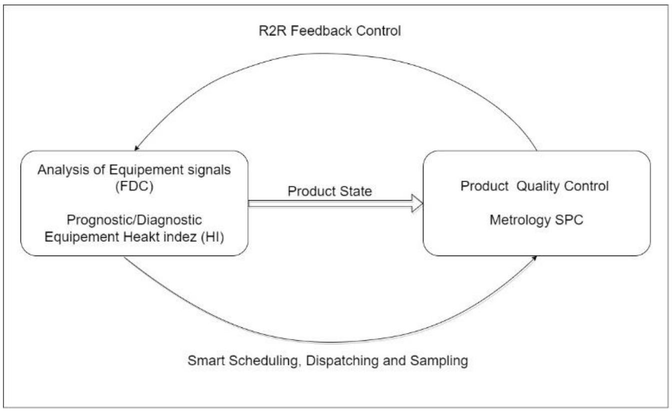

R2R Feedback Control
Analysis of Equipement signals (FDC)
Prognostic/Diagnostic Equipement Heakt indez (HI)
Product State
Product Quality Control
Metrology SPC
Smart Scheduling, Dispatching and Sampling

Figure 2.5: Advanced control loop

- From a process perspective, equipment deterioration modeling gives information about the equipment's qualification level. This information decides whether equipment is qualified to perform production tasks and which equipment becomes less qualified to perform operations. As a result, the factory can devise a productive and proactive production scheduling and dispatch system. With the help of this approach, dynamic dispatching and scheduling will play an important role in lowering variability in virtually all automated and advanced production factories, particularly in the semiconductor industry.
- From the equipment perspective, equipment deterioration modelling provides a diagnostic and prognostic maintenance plan to have a minimal maintenance cost and the minimum probability of unexpected equipment breakdown. Having fewer equipment downtime increases equipment utilization and production yield directly.
- From a product quality control point of view, equipment degradation modeling indicates equipment health and suggests readjusting or resetting process parameters using Run-to-Run system control, such as EWMA control, to avoid product failures. This feedback method not only reduces product waste and total manufacturing costs, but it also indirectly improves customer satisfaction and market share.

The simulation model proposed in this thesis is intended to describe this type of loop, highlighting the typical difficulties of making a good estimation of the level of degradation in the semiconductor equipment, due to the complexity of resuming and gathering enough feedback information about the health and infinite dynamics of semiconductor processes. The CBM performed over quality control will be integrated as tool to independently recognize deficiencies of the system and cover it with appropriate intervention unseen by the prognostic analysis and vice versa.

2.9. Enabling technologies and compatibility of PdM in semiconductors process

Since the APC approach described in the section 2.8 rely on resource reliability and degradation models obtained from field data, learning technologies and cognitive computing methods are relevant for the production quality scope. In the recent years, approaches for intelligent data analysis and classification have been presented in order to forecast machine and process behaviour and to give problem diagnostics based on predictor factors. A comprehensive review of these approaches can be found in [52]. There exist several recent techniques to deal with this issue. The most important include Decision and Regression Trees, Classification Rules, Fuzzy Models, Genetic Algorithms, Bayesian Networks, Artificial Neural Networks. Failure detection and classification are, in general, well established and accomplished nowadays [62]. Some strategies are already in use in the semiconductor industry, despite the fact that the level of complexities and the number of executed stages may pose a challenge to predictive methodologies. Various prediction algorithms are used in semiconductor equipment, a few of them are given below:

- Ion Implantation Tool
- Dry Etch
- Photolithography
- Chemical Vapor Deposition (CVD) Tool

[63] has fed an Artificial Neural Network with FDC dataset to detect semiconductors machine outliers. [64] describe three machine learning methodologies to respectively predict time to failure (TTF), the health state of the etching machine and TFF intervals of an equipment.

[19]] provide a predictive modelling strategy for intelligent maintenance in semiconductor manufacturing processes, based on machine in-situ sensor performance as well as product quality data. After that, they use self-organizing maps to discretize continuous data into discrete values, greatly reducing the computing cost of the Bayesian network learning process, which can uncover stochastic dependencies between process parameters and

product quality. This strategy, which differs from standard methods based only on inspection data, allows for more proactive product quality prediction.

Therefore, All the various methodologies it can be understood how PdM topics are so different one from each other and every new PdM problem should be studied separately with a customized solution.

However, it is not always feasible to assess the status of the processes with complete accuracy, both for a severe data reading issue and for a historical deficiency of sufficient observations to create a trustworthy statistical model: This is because much fewer maintenance actions were made than there were wafers to measure[20].

Therefore, wafer manufacturing facilities around the world have examined methods and techniques for the increase the yield production of products supplied, the increase of machine uptime, the reducing cycle time.

The main problems tied to data feedback are the following [19], [20]:

- **high dimensionality**. hundreds of input variables are available making the regression problem computationally expensive and difficult to solve.
- **Fragmented data and disconnected information** between maintenance Hundreds/thousands of products are run on the same machine, with different tool settings (called recipes);
- **Data not easily obtainable except by invasive diagnostic**
- **Missing data due to production line efficiency constraints** that do not allow every measurement to be taken at every step of every wafer
- **time series input data**. In many semiconductor modelling scenarios, the estimate of a scalar output from one or more time series is necessary. Usually, to address such problems, a fixed number of characteristics from time series (such as statistical moments) are extracted, which leads to data loss and worse prediction models.
- **multi processes modelling.** There are several sequential activities involved in the manufacture of semiconductors, and the quality characteristics of a given wafer depend on the whole processing process rather than just the final step before measurement.
- **Limited and reliable historical data** also due to frequent parameter set up and reset
- **Limited information knowledge about the control of certain processes** not yet fully perceptible such as contamination phenomenon in chamber tool processes

3 Recap of response surface methodology (RSM)

3.1. Performance evaluation models

Multi-stage production and transfer lines are made up of a series of equipment that may perform specialized operations on raw materials at a set rate. The reliability of the process carried out by the machines has a significant impact on MMS performance. Given the random nature of the phenomena involved in the behaviour of such structures, a stochastic method is required to create adequate models to assist manufacturers in implementing the best strategy for future customer requests. [64] provide a classification of existing performance evaluation models as well as a review of them. The following are the three most important tools: Models based on Queuing Networks, Markov Chains, and Simulation Models

Simulation Models as Discrete Event Simulation (DES) are widely used by manufacturing companies since they can reach high levels of detail during their design. However, as it is already explained in section 1.2, they are generally time consuming, and this can be a problem during the design phase, when large number of alternatives need to be evaluated. Furthermore, a simulation is an experiment, and must be repeated to have a statistically reliable result. Anyway, Simulation models could bring out a range of outcomes closest to the reality.

Queuing Networks can give an exact solution but under the constraint of restrictive hypothesis.

Analytical models based on Markov Chains can be placed in the middle between the approaches mentioned above and can give accurate result with less restrictions than Queuing Network models with more realistic hypothesis. These models are differentiated based on the assumption of the flow parts and the machine processing times.

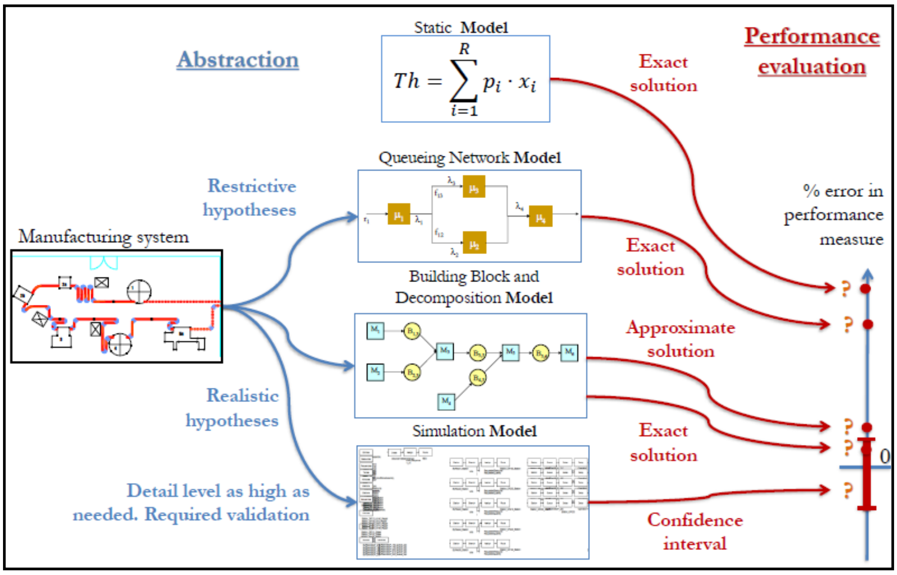


Figure 3.1: Manufacturing system model comparison [65]

3.2. Performance evaluation for two-machine lines through metamodels

Due to a global increase in competitiveness in manufacturing, companies strive to increase the effectiveness of their manufacturing systems. The new industrial revolution, Industry 4.0, is a consequence in motion to aid in creating improved manufacturing systems. A common tool within Industry 4.0 is simulation, where one could simulate changes in a virtual representation of a real-world system.
Discrete Event Simulation (DES) is a tool that has been widely adopted within industries to test manufacturing system changes virtually before implementing them physically.

Some of the most common simulation methods are discrete event, continuous and agent based [66], [67]. Systems can be categorised as discrete or continuous, and the agent-based approach's simulation method can be used to both types of systems. Because DES uses an event-driven simulation approach, system changes are triggered by events rather than the passage of time as in continuous systems. It is easy to distinguish between these two systems by referring to the discrete systems as a bank. The client enters the bank, waits for their turn, gets assisted, and then leaves. The customer determines this mechanism, which triggers a change when the customer's status in the bank changes. Water flowing from a dam can be thought of as moving continuously and time-dependently in a continuous system. ABS is precisely what it sounds like: it is based on models referred to as agents. In order to evaluate the rest of the system, the agents communicate with one another. A series of rules determines which agent to use, and they must be followed in that order. Nonetheless, model dynamics are unable to predict a certain level of autonomy. Because the agents have intelligence, awareness, memory, and contextual awareness, this is the case. The background study [67] concluded that the DES was the most suitable method for this project and therefore of big interest. The key features of DES simulation are stated below .

- Predefined start and end points.
- An event-driven simulation method.
- Events occur instantaneously and therefore the time step in between processes are zero.
- The sequence of events are stored in an event-queue to be executed in the correct order that is determined by the user.

Workload balancing, resource allocation, capacity planning, layout design, inventory level assessment, supply chain management, and other operational and tactical difficulties that emerge in the manufacturing realm have long been resolved using discrete event simulation (DES)[68]. DES can be used to study the current production line, investigate prospective improvements, and compare different options while simulating the real production line with minimal assumptions. DES is particularly well suited for modelling manufacturing systems as DES can explicitly model the variation within manufacturing systems using probability distributions. SM has a lot of limitations, including computing time, black box properties, and the inability to do optimization analysis directly, while being frequently used to cost-effectively analyse the behaviour of complex industrial systems without interrupting production. Metamodeling combines the benefits of SM and design of experiments by reaching optimal solutions in a short amount of time and the capacity to analyse complicated systems with a deterministic outcome in an explicit form [69].

Generally, in literature a production line composed by two machines and one buffer is called two-machine one buffer line (2M1B) or Building Block (BB). Various authors have developed performance evaluation models for two-machine lines. [70] analyses two machine lines considering multiple failure modes of the machines and finite buffer capacity. [71] develops an analytical model for deterministic asynchronous two-machine line ruled by some threshold-based control policy. These and other models [11] take advantage of the assumption that a machine can be represented in terms of a specific Markov chain based on the buffer level. Regardless, there are rare circumstances in which an analytical model is built with such sophisticated dynamics that it may include Maintenance, production, and quality interactions in a unique thorough framework in the performance evaluation. Furthermore, if the Natural Extend is the analysis of longer production lines, no exact analytical models are accessible for such a system due to mathematical tractability. Thus, Discrete event simulations improved by Metamodels is commonly studied by researches to get right performance and predictions future outcomes.

One of the most useful tools in modern business, optimization enables decision-makers to more assertively allocate their resources and deal with challenging industrial challenges. Additionally, the interdependence and high stochastic levels of production systems make the use of optimization and simulation tools to address challenges necessary. Although simulation optimization is a powerful tool, it might take a while to discover a workable solution, which limits its use in routine tasks. In the optimization process, metamodels can be utilized as an alternative to simulation models.

It is called simulation model when the mathematical model of a system is studied using simulation. By running the simulation model for a predetermined amount of time, the behavior of the system at particular input variable values is assessed. An experiment or series of experiments known as a simulation experiment include making major changes to the input variables of a simulation model in order to track and pinpoint the reasons for changes in the output variable. The simulation experiment may become computationally expensive when the number of input variables is high and the simulation model is intricate. In addition to the high computational cost, choosing sub-optimal input variable values results in an even larger cost. Selecting the best input variable values for a simulation is the act of looking through all of the options without explicitly evaluating each one. Its goal is to spend as little money as possible while getting the most information possible from a simulation experiment.

In general, the simulation optimization problems can be stated as min $\theta \in \Theta$ f (θ) where θ can be a single variable or a p-dimensional vector of all the decision variables, and Θ is the feasible region. For simulation optimization problems, we do not have much knowledge on the structure of f (θ) and the analytical expression of f (θ) cannot be obtained or may not even exist.

Therefore, the objective function must be estimated based on the outputs of simulation runs, such as f (θ) = E[L(θ, ω)] [72]. There are many practical problems related to simulation optimization in the real world [30], [73]–[75].

Unlike other optimization problems, like linear programming and mixed integer linear programming, there are some difficulties in solving optimization problems:

- It is not present an analytical expression of the objective function.
- The randomness of the simulation model entails different simulation replication run
- In many cases simulation is time consuming

In the Fig.3.2 below a classification of the main metamodels is addressed. According to the underlying structure of the decision variables, two cluster of simulation optimization are identified:

- Continuous decision variable, dedicated for problems with continuous, uncountable, and infinite variables.
- Discrete decision variables, which address optimization issues when the viable range of the input variables is finite or countably infinite. With this kind of variable, the techniques listed in the first bullet point might not work.

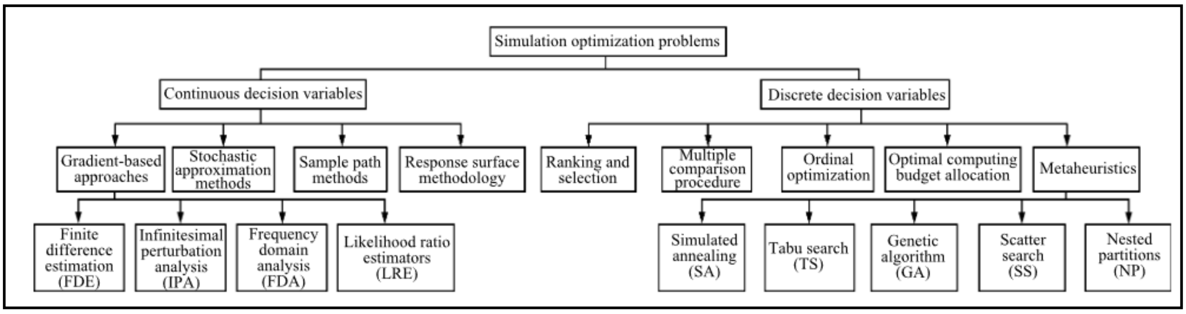

Simulation optimization problems
Continuous decision variables
Discrete decision variables
Gradient-based approaches
Stochastic approximation methods
Sample path methods
Response surface methodology
Ranking and selection
Multiple comparison procedure
Ordinal optimization
Optimal computing budget allocation
Metaheuristics
Finite difference estimation (FDE)
Infinitesimal perturbation analysis (IPA)
Frequency domain analysis (FDA)
Likelihood ratio estimators (LRE)
Simulated annealing (SA)
Tabu search (TS)
Genetic algorithm (GA)
Scatter search (SS)
Nested partitions (NP)

Figure 3.2: Metamodel taxonomy

3.3. Response surface methodology as metaheuristic tool

Process optimization normally involves the combination of mathematical and statistical techniques which can be approached by distinct ways. Despite the fact that different methods can be found in the literature, the response surface methodology raised as one of the most effective ways for performing process optimization, by combining design and analysis of experiments, modelling techniques, and optimization methods. Figure shows how RSM may be viewed as a combination of these components. This intersection of procedures implies that researchers should be very mindful in each one of the three steps involving RSM. Without this caution, this methodology will certainly fail, does not produce the expected and possible desired results. In Fig. 3.3 a Road map is delineated in [76] for carrying the RSM analysis.

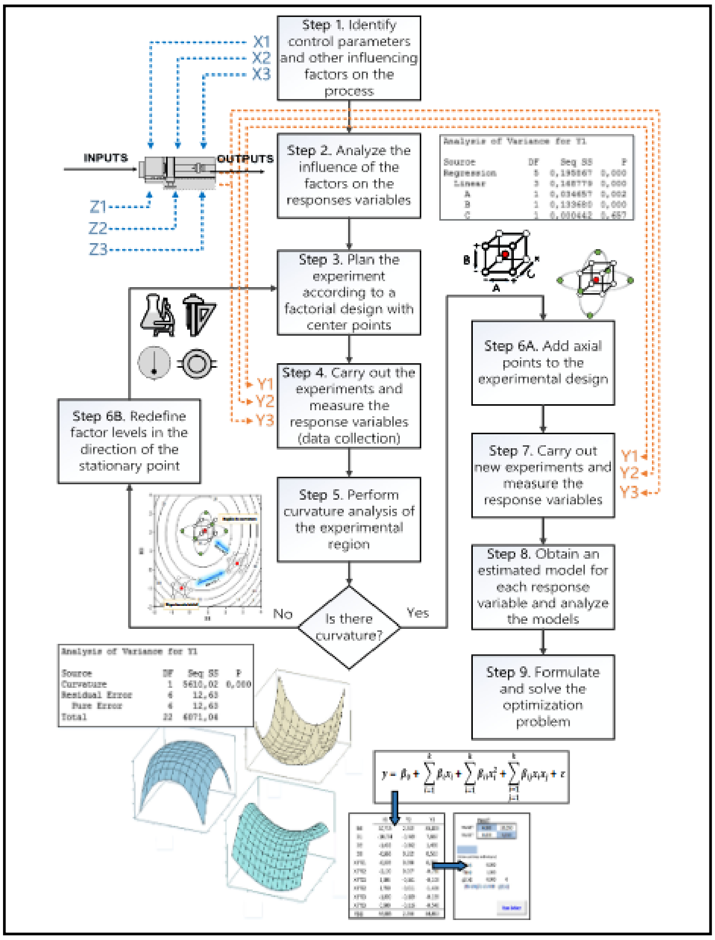

Step 1. Identify control parameters and other influencing factors on the process
X1
X2
X3
INPUTS
OUTPUTS
Z1
Z2
Z3
Step 2. Analyze the influence of the factors on the responses variables
Analysis of Variance for Y1
Source DF Seq SS P
Regression 5 0,195867 0,000
Linear 3 0,148778 0,000
A 1 0,034657 0,002
B 1 0,133680 0,000
C 1 0,000442 0,657
Step 3. Plan the experiment according to a factorial design with center points
B
A
C
Step 4. Carry out the experiments and measure the response variables (data collection)
Y1
Y2
Y3
Step 6A. Add axial points to the experimental design
Step 6B. Redefine factor levels in the direction of the stationary point
Step 5. Perform curvature analysis of the experimental region
Step 7. Carry out new experiments and measure the response variables
Y1
Y2
Y3
No
Is there curvature?
Yes
Step 8. Obtain an estimated model for each response variable and analyze the models
Analysis of Variance for Y1
Source DF Seq SS P
Curvature 1 5610,02 0,000
Residual Error 4 12,63
Pure Error 4 12,63
Total 22 6071,04
Step 9. Formulate and solve the optimization problem
$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$

Figure 3.3: Road map for an efficient conduction of response surface methodology

1) The first step is to determine the variables governing the process studied, which are the control parameters also called factors.
2) Explanatory experiment design is the most popular strategy for the next phase. To explore a big number of variables using a small number of trials, it includes setting up a limited number of runs, such as fractional or factorial. Finding out how the variables impact the process under examination is the aim of this phase. Since it enables identifying which

factors are genuinely significant and their individual influences on response variables among the factors investigated, ANOVA (Analysis of variance) is the statistical method that is most frequently utilized in RSM research for this purpose.

3) The third step is to plan data collections according to an experimental design. For this purpose, compose the design of experiments (DOE), central composite design (CCD), Box-Behnken design (BBD), and Taguchi designs are the most usual methods.
4) The responses variables are measured and collected
5) Determine if the experimental region is curved; if not, add axial points to the design and carry out fresh trials.

   Utilize the data gathered to create mathematical models that accurately depict the process under study in relation to its control parameter, creating a second-order model using the ordinary least squares method.

6) The last step of RSM is to find the set of control parameters that improve the process by applying an optimization method to minimize or maximize the functions modelled in Step 6.

4 Problem statement

4.1. Integrated framework for manufacturing systems with deteriorating equipment design and analysis

Manufacturing systems in semiconductor industry involves multiple components, stations and operations through which the final wafer is produced. Therefore, it is characterized by more complexity in handling their parameters and dynamics to coordinate interactions among all the stages.

In this thesis it is intended to model an equipment deterioration of Batch Manufacturing Process that are pervasive modes in today's semiconductor fabrication. As a matter of fact, degradation of a component/system is one of the major factors that cause defective product output.

Moreover, as the part advance through all the stages, product quality variations are introduced and propagates. The final product's quality attributes are calculated by summing together all mistakes committed at any stage. The problem with model-based process control has been examined in terms of how measurement data up to any given operation, the history of prior control actions, and the model are utilized to carefully choose controlled policy over process parameters, so slightly reducing outgoing quality defects.

Therefore, two particular control visibility aspects of the system are meant to be examined with this goal in mind: The first is implemented locally to keep an eye on the equipment's health and guard against serious malfunctions that might endanger system production and, indirectly, the quality and precision of the final output. On the other hand, a quality control policy may be remotely monitored at the downstream level to identify indirect misleading behaviour of the process through quality deviations from the typical state of the product. This joint approach is considered in this work since it is still impossible to derive from the equipment's state condition a general rule of all the product quality characteristics. Because of this, even with the advent of predictive maintenance, inspection and metrology techniques remain crucial in the semiconductor production process.

Considering the first control mentioned above, different enabling tools are present in literature [19], [21], [63], [64]. These tools have been taking the advantage of digital architectures, which enable more efficient analyses in equipment deterioration modelling. The semiconductor manufacturer benefits from such booming growth of data as an important key for APC solutions to prognose equipment deterioration and diagnose failure causes.

However, there rarely exists condition-based maintenance, which utilizes machine conditions to schedule maintenance, and almost no truly predictive maintenance that assesses remaining useful lives of machines and plans maintenance actions proactively. Currently, the majority of maintenance operations in the semiconductor industry are still based on either historical reliability of fabrication equipment, or on diagnostic information from equipment performance signatures extracted from in-situ sensors. Such a fragmented, "diagnosis-centered" approach leads to mostly preventive maintenance along with reactive maintenance policies that use neither abundant product quality, equipment condition, equipment reliability information, nor the temporal dynamics inside that information in order to anticipate future events in the system and thus facilitate a more proactive maintenance policy.

The semiconductor manufacturing industry has been unable to implement a more pro-active, "prediction-centered" maintenance strategy based on the available on-line sensing, quality control, and reliability data collected throughout the shopfloor due to a number of research difficulties. The following issues impede this step of prediction for semiconductor operations:

First, due to the high system complexity, it is almost impossible to observe any analytical or deterministic phenomena in the fab. Inherent stochastic nature of a semiconductor fabrication, in which production and maintenance operations are constantly interacting, needs to be modelled and then used to predict equipment behaviour and facilitate a proactive maintenance.

Second, it is difficult to observe always the state of the equipment. It is a key element in the semiconductor industry that enables CBM and PdM. However, due to the complexity of the process dynamics, it is difficult to consistently and cost-effectively examine this indication using existing monitoring approaches.

Third, the complex interaction between equipment degradation, product quality, maintenance operations and production process is another challenge. The present fragmented and individually regarded maintenance, production, and inspection databases must be taken into consideration at the same time in order to achieve fully proactive maintenance. Collaboration and a network connecting maintenance, manufacturing, and quality control staff are necessary for this.

Finally, the batch process data collected from the equipment are commonly referred to the FDC data in the semiconductor fabrication industry. The extraction of the health index (HI) or degradation level (DL) from a huge and diverse data size is the first problem in the equipment deterioration modeling of FDC data. Another issue present and still not assessed in the simulation model is multiple recipe contexts in semiconductor fabrication. For instance, in semiconductor manufacturing process, almost all processes (e.g. litography etching and deposition) are carried out with different process recipes [7], [19].

On the quality point of view, the objective of the inspection station are both to identify as soon as possible the OOC through Run-to-Run process EWMA control chart and monitor a condition-based point of quality to monitor misbehaving level of defectiveness and make appropriate intervention to the upstream process. In this way it is possible to restore the system even if the health condition of the equipment is still considered good, giving the priority on the product specific claims rather than the operating time. However, the main problematic of remote measurement stations is that they only take into account the decrease of quality fluctuations, disregarding upstream system dynamics such as the delay between the production of a faulty item and its inspection. The number of non-conforming products manufactured during this time lag will decrease the system yield. Therefore, it is important to evaluate the quality problem both from the process and the system point of view.

Consequently, the two main issues are to identify potential information about the behavioural changes of FDCs and to create an effective model for estimating equipment degradation. This model will be employed to forecast the state of the equipment and the underlying cause of failure. This method is strengthened by the inspection control quality process (SPC) in order to find quality deviations that the machine controller is unable to avoid on time.

This joint control policy can be effective or not according to the appropriate Manufacturing system design which may foster the feedback advice (shown in Fig.4.1) of both sides with a certain tradeoff. To this reason, the quality perspective and machine perspective interactions must be assessed in order to see what kind of control policy is predominant with certain configuration of the system, and whether there is a trade-off between sensing cost implementation and system performance. Simulation models and metamodels plays an important role as performance evaluation tool to describe this possible real behaviour of the system as close as possible and provide an important solution for decision making. The simulation aims to highlight the productivity performance of the line both from the productivity indicators point of view (such as total throughput, delay of feedback, number of repair) and quality KPIs (yield, effective throughput, defective throughput, scrap rate).

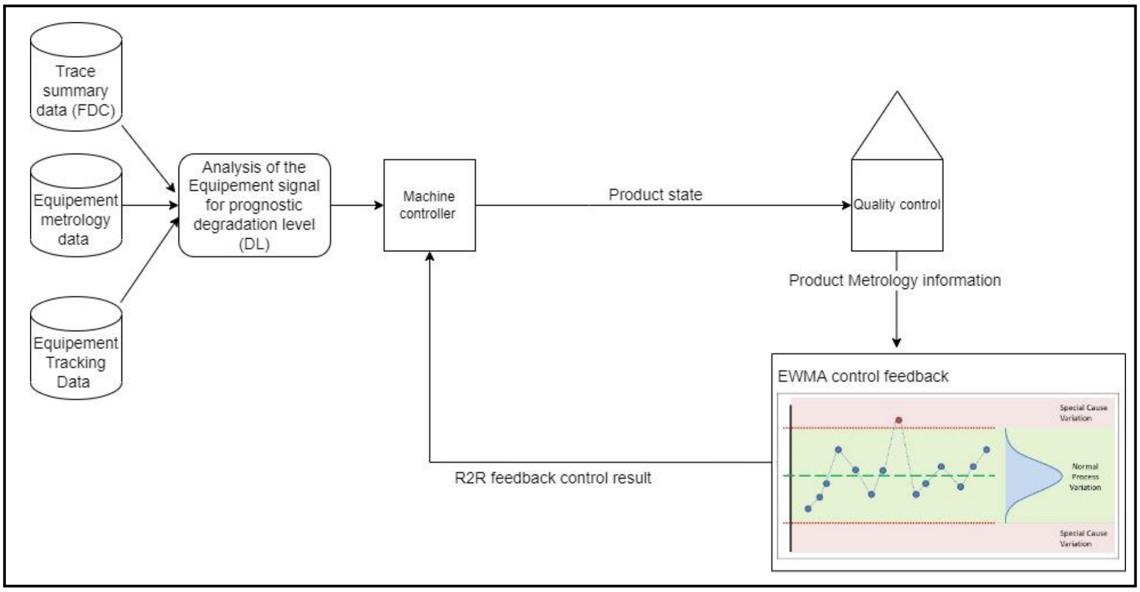


Figure 4.1: Scheme that summarizes the information flow as enabling tool for the decision making

4.2. The outline of the method

This thesis takes into account the issue covered in the preceding section. So, simulation is specifically designed to reflect the effects of both inspection CBM and predictive maintenance for the upstream process. The Descrete Event Simulation developed will deserve as a base for a further simulation-optimization approach, in which computer simulation are combined with optimization technique to solve problems that are analytically intractable. The solution approach combines mathematical modeling, simulation techniques, design of experiment and response surface methodology with the aim to replace the complex model with an approximated model that we can optimize, leading to the optimal values of the control parameters. This methodology imitates the stochastic and complex behaviour of the production system and has successfully solved many complex optimal control problems[36], [77], [78] .

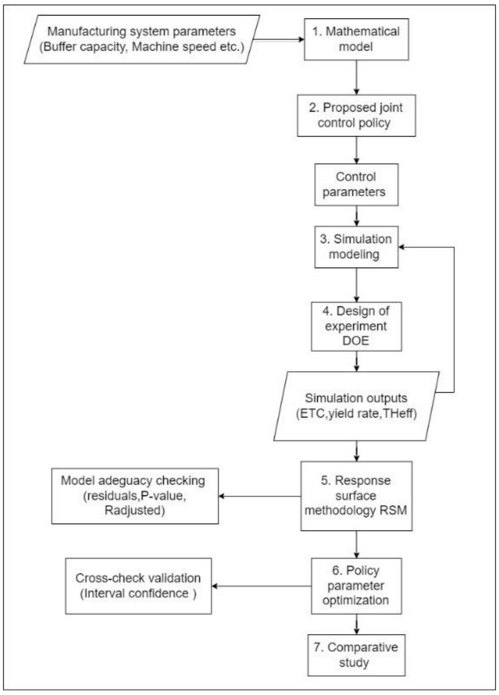

Manufacturing system parameters (Buffer capacity, Machine speed etc.)
1. Mathematical model
2. Proposed joint control policy
Control parameters
3. Simulation modeling
4. Design of experiment DOE
Simulation outputs (ETC,yield rate,THeff)
5. Response surface methodology RSM
Model adeguacy checking (residuals,P-value, Radjusted)
6. Policy parameter optimization
Cross-check validation (Interval confidence)
7. Comparative study

Figure 4.2: Road map for an efficient conduction of response surface methodology

The resolution approach shown in Fig 4.1 consists of the following systematic steps:

1) *Mathematical modelling*: This step consists in the analytical formulation of the production system under study as detailed section 6. This step provides a detailed model of the system dynamics, the objective function to be minimized, the definition of the decision variables and the problem constraint.
2) *Determination of the joint control policy*: Based on literature examples, a joint control policy is proposed in section 6. The control policies are characterized respectively by control parameters for the quality sampling(s) with the usage of a deviation limit of EWMA function (CBM) and the level of degradation limit to intervene directly on the machine (PM). The policy faces random events like failures, repairs and effects of deterioration.
3) *Simulation model*: The mathematical model is transformed into a discrete-continuous simulation model following the logic of section 6.2.
4) *Design of experiment:* this step uses the outputs of the simulation model to conduct a factorial experimental design (3 level and 2 factors) and inscribed experimental design jointly. The experiment is used to determine with a minimum number of simulations runs the main factors, interactions and quadratic effects of the control parameters that significantly affect the simulation model outputs and must be considered in the optimization step.
5) *Response surface methodology:* Once significant factor is identified, we determine second-order regression metamodels, based on the response surface methodology (RSM), for the expected total cost (ETC), the yield rate and the effective throughput. The quadratic regression function for both the output item are formulated as follow:

$$ETC(PM, CBM) = \gamma_0 + \gamma_1 CBM + \gamma_2 PM + \gamma_3 CBM\ PM + \gamma_4 PM^2 + \gamma_5\, CBM^2 + \varepsilon \qquad (4.1)$$

$$THeff(PM, CBM) = \gamma_0 + \gamma_1 CBM + \gamma_2 PM + \gamma_3 CBM\ PM + \gamma_4 PM^2 + \gamma_8\, CBM^2 + \varepsilon \qquad (4.2)$$

$$Yield(PM, CBM) = \gamma_0 + \gamma_1 CBM + \gamma_2 PM + \gamma_3 CBM\ PM + \gamma_4 PM^2 + \gamma_5\, CBM^2 + \varepsilon \qquad (4.3)$$

Where $\gamma_i, i\epsilon(1,5)$ are regression coefficient estimated for each regression and ε is a random error component that incorporates all other resources of variability.

The adequacy of the regression metamodels is checked in the region of the optimal solution with the adjusted coefficient of determination R-squared that should be close to one for these expressions. Also, a complete examination of residuals is performed to ensure the normality assumption and their homogeneity

6) Parameter optimization: Once we obtained the regression models, the quadratic functions replace the unsolvable model with an approximated one. The two functions can be optimized through non-linear constrained optimization techniques such as Genetic algorithm. In this case the MATLAB software was used.
7) Optimization models are formulated in section 8. The first decision making optimization is aimed to maximize the throughput effective. The second optimization focuses on the maximization of the yield rate. Finally, the minimization of an objective cost function both dependent by ETC and THeff functions is addressed. Upon the optimization, the optimal solution is cross- checked with extra solution runs to define a confidence interval for the expected total cost
8) Different configurations of the manufacturing system are assessed in section 8. This comparative analysis is performed in order to prove that the initial parameter condition of the system strongly affects the quality and maintenance policy decision making.

Regression metamodels have been a successful alternative to determine an optimal solution for complex systems. The sequential procedure of DOE, regression modelling and constrained optimization must be conducted in an appropriate range for the control parameters to fully explore the entire admissible control domain and determine a close approximation of the optimal solution.

5 Case description

5.1. Case study

The model application to an industrial example is described in this chapter. First and foremost, a general overview of semiconductor manufacturing with a focus on wafer fabrication is provided. Secondly Problem of degradation and training model for APC machine process control is addressed. Finally, the model is applied to the specific case to demonstrate how the combined application of SQC techniques and PM methods for achieving higher product quality and more effective use of resources is affected by the design of system configuration.

5.2. The fabrication of a semiconductor device

Wafer fabrication is the process of building an integrated circuit on raw silicon wafers by layering complicated designs together. Every layer of the wafer is distinguished by the recurrence of a sequence of processes. Every wafer has a number of dies, each of which is an Integrated Circuit (IC).
The manufacturing phase of an integrated circuit can be divided into two steps. The first, wafer fabrication, is the extremely sophisticated and intricate process of manufacturing the silicon chip. The second, assembly, is the highly precise and automated process of packaging the die. Those two phases are commonly known as “Front-End” and “Back-End”. They include two test steps: wafer probing and final test[8].

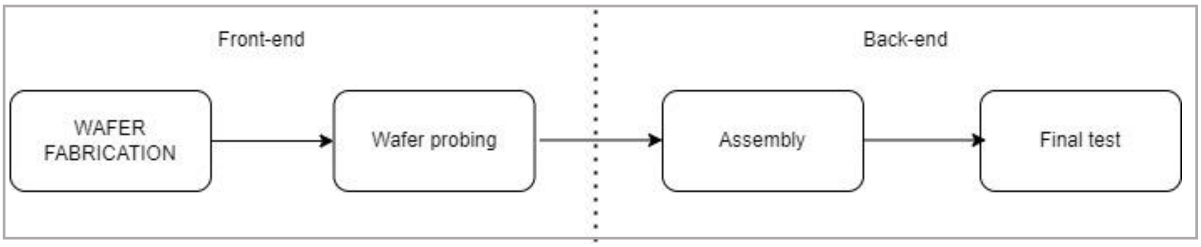


Figure 5.1: Manufacturing flow chart of an integrated circuit

5.3. Front-end

On each wafer, identical integrated circuits, or "dies," are built using a multistep process. On the wafer, each step either adds a new layer or modifies an existing one. These layers are the building blocks of individual electrical circuits. The fabrication process for a semiconductor wafer is composed of sequential main macro-steps (Fig.5.2) but also many other intermediate steps, such as inspection, cleaning, and other minor operations. At various phases of the procedure, some of them are repeated numerous times. The main steps to manufacture a single wafer layer are: [[7],[8]]

- **Oxidation or Passivation**: a thin silicon oxide layer is grown onto the wafer surface, working as an insulator.

- **Lithography**: In order to create the features that control the operation of the microelectronic device, a wafer coated by a photoresist layer is subjected to a laser that transfers a pattern from a mask onto the wafer surface. To create a nanometric circuit on the wafer surface, a laser beam is first formed by a mask and then scalarly reduced by lenses. The photoresist is then chemically eliminated.

- **Etching:** This operation removes a thin film material. There are two different methods: wet (using a liquid or soluble compound) or dry (using a gaseous compound like oxygen or chlorine).

- **Diffusion:** During this process, dopants are incorporated into the material, or a thin oxide layer is grown on the wafer. Doping gazes penetrate or interact with silicon on wafers placed in a high-temperature furnace (up to 1200 ° C), which causes a silicon oxide layer to form.

- **Ionic implantation**: It allows to introduce a dopant at a given depth into the material using a high energy electron beam

- **Metal deposition**: To enable the development of electrical connections between the various integrated circuit cells and the outside, a layer of metal is deposition onto the wafer surface. Evaporation or plasma sputtering might be used to carry out this operation.

- **Chemical-mechanical polishing**: the surface is polished to remove the metal excess and the eventual remaining photoresist, and also to obtain a perfect plane surface to repeat all steps for the next layer

- **Back-lap**: It's the last step of wafer fabrication. Wafer thickness is reduced (for microcontroller chips, thickness is reduced from 650 to 380 microns), and sometimes a thin gold layer is deposited on the back of the wafer

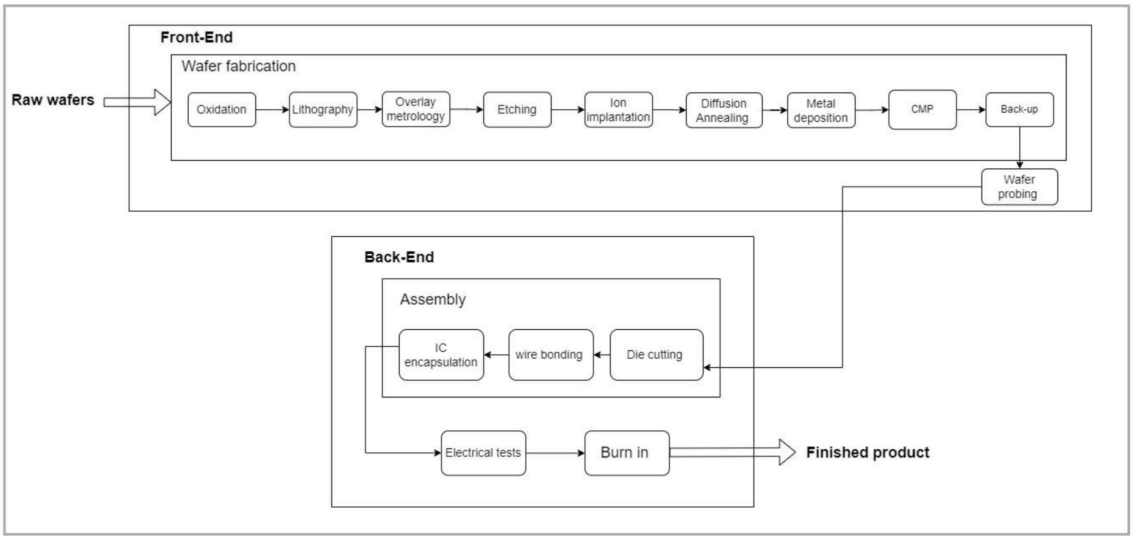


Figure 5.2: Wafer production steps

The chips are put through a series of electrical tests using specialized micro-probes connected to the dice on the wafer, ensuring their operation and classifying them based on their electrical characteristics. At the start of the assembly step, when the dice are separated using a cutting technique, the defective dice are marked so that they are discarded.

5.4. Diffusion and Ionic Implantation Process

The silicon chip is initially a piece of a raw wafer, a slice of silicon that is very thin (about 650 microns). Wafers commonly have diameters of 125, 150, or 200 mm (5, 6 or 8 inches). However, the major electrical characteristic of raw, pure silicon is that it is an isolating substance. As a result, some silicon properties must be changed using a carefully controlled technique. This is accomplished by "doping" the silicon. The silicon lattice is intentionally filled with dopants (or doping atoms), which alter the properties of the material in certain places. They are classified into "N" and "P" categories, which stand for the negative and positive ions they contain, respectively. The most widely employed dopants are phosphorous, arsenic (N type), and boron (P type), which are all used to obtain the necessary properties. Semiconductors manufacturers purchase wafers predoped with N or P impurities to an impurity level of.1 ppm (one doping atom per ten million atoms of silicon). The silicon can be doped in two different methods. The wafer is initially placed into a furnace. The silicon surface is subsequently impregnated with doping gases. Diffusion is a manufacturing process that includes this step (the other part being the oxide growth). Ionic implantation is the name of the second doping method for silicon. In this instance, an electron beam is used to inject doping atoms into the silicon. Ionic implantation, in contrast to diffusion, enables the placement of atoms at a specific depth inside silicon and, in general, provides for a greater control of all the process' key variables. Ionic implantation process is simpler than diffusion process but more costly (ionic implanters are very expensive machines).

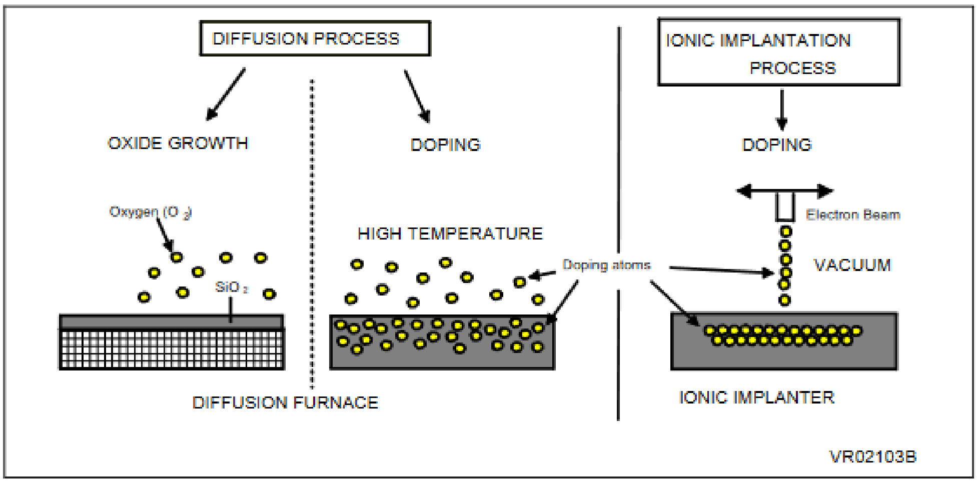


Figure 5.3: Diffusion process and Ionic implantation[8]

5.5. Photolithography

The process is performed by a machine called stepper that passes light through a mask, forming an image of the reticle pattern. The image is focused and reduced by a lens and projected onto the surface of a silicon wafer that is coated with a photosensitive material called photoresist. The wafer is shifted below the optical system after each exposure by precisely the size of the image field, and completely exposed step by step hence the name “wafer stepper” (Fig.5.4).

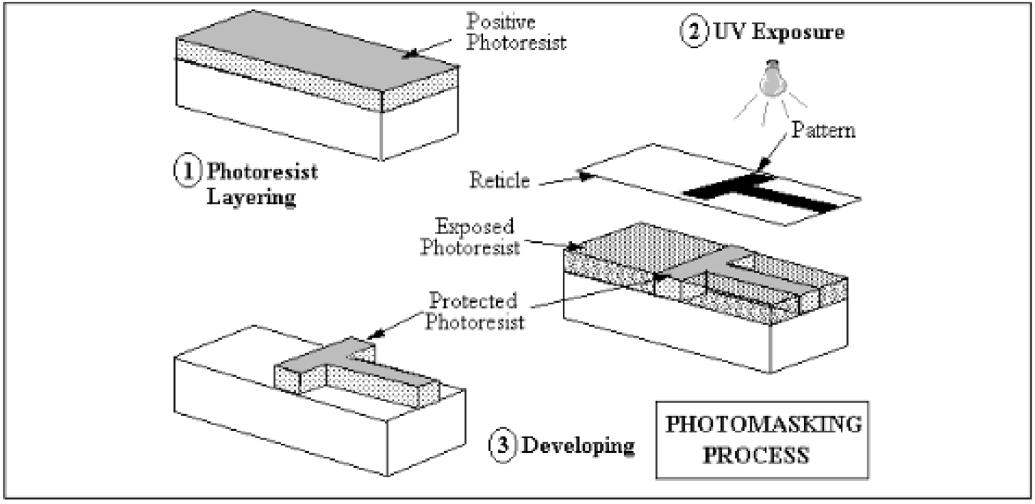


Figure 5.4: Diffusion process and Ionic implantation

The covered wafer is developed like photographic film after exposure in the stepper, which causes the photoresist to dissolve in certain spots depending on how much light those regions got during exposure. The next step is etching, which involves exposing the created wafer to acids or other chemicals. In the areas of the wafer that are no longer covered in the photoresist layer, the acid dissolves the silicon and removes it. Production is done in lots, and the stepper's setup settings and mask patterns are often the same for at least one of the lots.

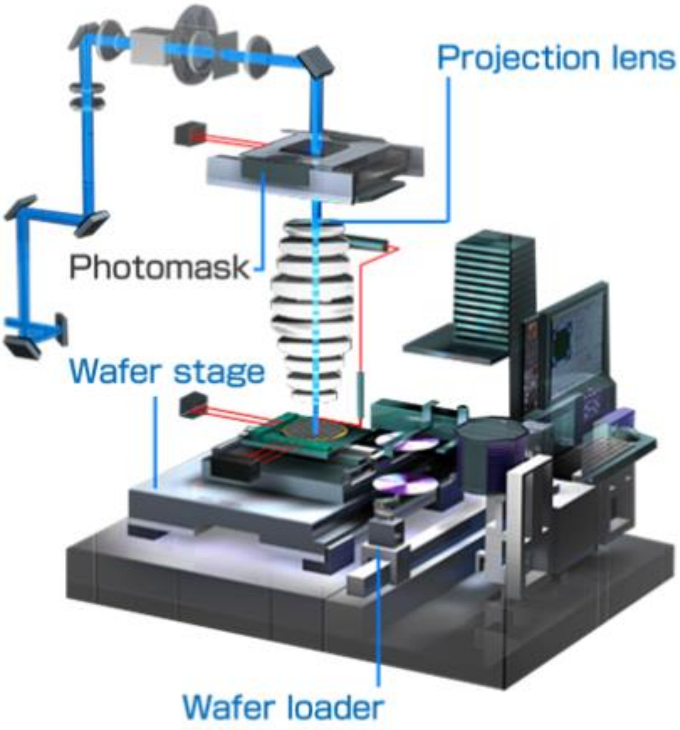


Figure 5.5: Real example of photomask lithography [7]

Photolithography, which accounts for 33 % of the expenses associated with wafer manufacture, is crucial to IC manufacturing because to the accuracy requirements. Numerous similar procedures are later carried out to build a whole semiconductor wafer, and each pattern transfer has a highly accurate location on the wafer surface. Overlay refers to how each layer lines up with the layer that came before it. A perfect overlay is essential for the quality of the devices that are manufactured because it enables appropriate electric current flow in the integrated circuit.

5.6. Metal deposition and Etching process

Metal deposition is used to put down a metal layer on the wafer surface. There are two ways to do that. The process shown on the graph below is called sputtering. It consists first in creating a plasma with argon ions. These ions bump into the target surface (composed of a metal, usually aluminium) and rip metal atoms from the target. Then, atoms are projected in all the directions and most of them condense on the substrate surface.

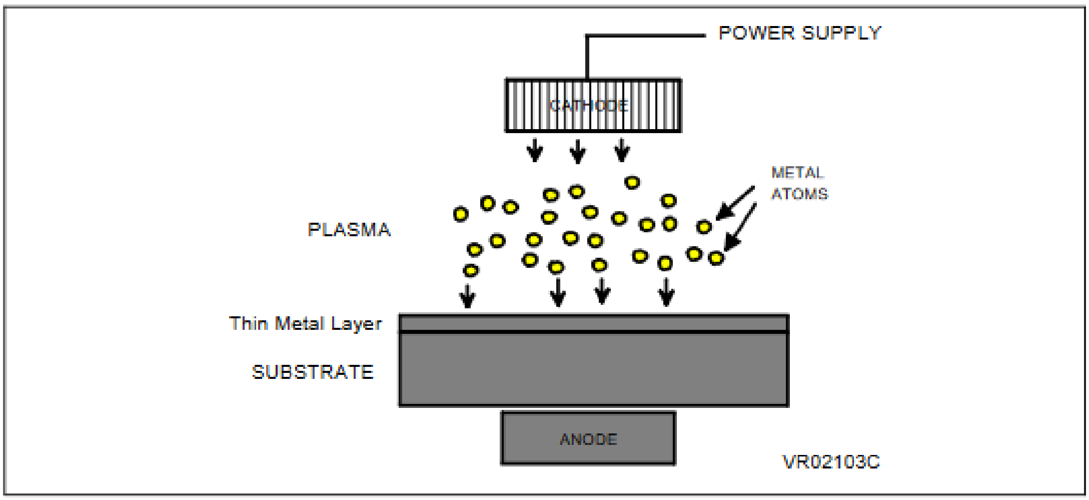


Figure 5.6: Metal Deposition Process[8]

The circuit layout produced during the photomasking process is etched onto a particular layer using the etching technique. The layer that has to be etched is typically deposited first, then the etching process begins. For instance, etching the poly layer produces the poly gates of a transistor. The connections made from aluminium after the aluminium layer was etched serve as a second illustration.

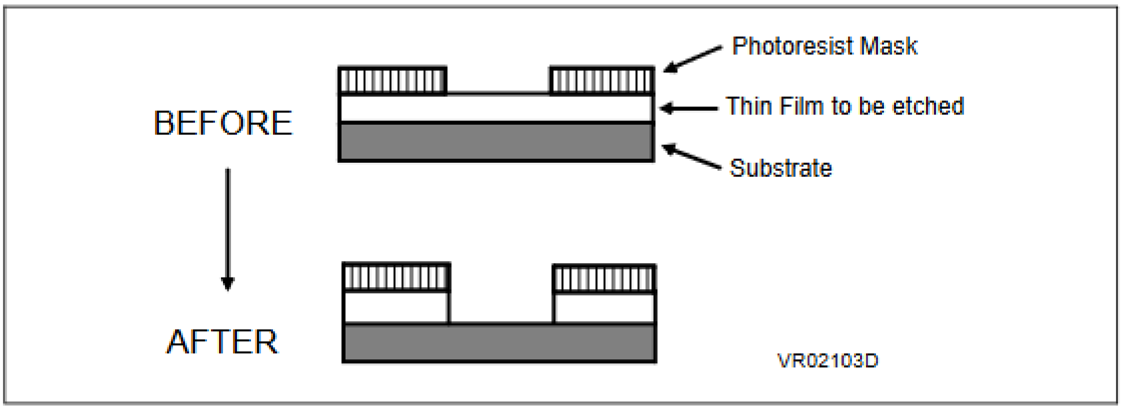


Figure 5.7: Metal Deposition Process

5.7. Problem Description

A sophisticated manufacturing system, like a semiconductor fabrication plant, often entails hundreds of production processes and a wide range of equipment. Millions of dollars in capital expenditures may be required for the manufacturing of a single tool or wafer scrap. A significant loss in productivity and revenue might come from equipment downtime. Furthermore, due to the complexity of the production process, downtime on a single tool might result in delays and idle time on several additional fabrication machines[19]. Therefore, maintenance is essential to keep tools running at their peak performance levels. Several research challenges have prevented the semiconductor manufacturing industry from achieving a more proactive, "prediction-centered" maintenance approach based on the available on-line sensing, quality control[53], [79], and reliability data collected across a fab[20] [80] [79] [19]: It is quite challenging to first detect any analytical or predictable events in the fab because to the tremendous system complexity. It is necessary to characterize the semiconductor fab's intrinsic stochastic nature in order to utilize it to anticipate equipment behaviour and enable preventive maintenance. Production and maintenance activities there are continually interacting. A problem with the equipment is the unobservable state. The most reliable degradation indicator in chamber tools, like diffusion etching and metal deposition, is particle counts, which is the key element enabling the CBM and PdM in the semiconductor industry[80]. This indicator, however, is hard to be cost-effectively and reliably observed using current monitoring techniques. On the other hand, the research in modelling particle counts using available process and product measurements did not give satisfactory results. The complex interactions between the production process, product quality, maintenance tasks, and equipment deterioration are challenging to manage. To properly implement proactive maintenance, the current disparate and separately considered production, inspection, and maintenance databases must be taken into account simultaneously. Collaboration and a network connecting maintenance, manufacturing, and quality control staff are necessary for this.

All of these issues are also handled for those processes operating before the overlay metrology, which is the primary inspection station. Numerous externally hampered events that regularly influence lithography and the oxidation process have a long-term impact on the system's health.

5.8. Organic and airborne contaminations

In Oxidation as explained below, possible contamination of the oxidate layer may occur significantly affecting the reactive properties of the film over the wafer. Organic contamination is transferred by the environment or by media to surfaces. Furthermore, organic contamination from previous processing steps remains on the surface. Therefore, also succeeding step of metrology inspection (CVD, Metal Diffusion, Ion implantation) may impact on the quality result, since manufacturing fabrication is characterised by a ring loop. It is frequently unknown how organics particle affects surfaces on nanostructures and gadgets, as well as whether cleaning procedures can remove any leftover contaminants. There are many different organic substances and potential sources. As a result, a major metrology difficulty is reliable detection and categorization. [19], [81].

As well, lithography has partial knowledgeable process misleading behaviours. This is crucial because the ray projection needs to be very precise. A small deviation could mean the difference between a working chip and a defect chip. Therefore, it is very important that the lenses are used correctly for the projection. The overlay metrology inspection measures how well the projection is done to adjust the lenses. This station is fundamental for the corrections and control of the overall quality of the process. More in detail, in the photolithography process where many printed layers need to be aligned one onto each other with sub-nanometric precision, the inspection station is considered the system bottleneck and the major capital effort is implemented here to improve the efficiency of the system. A new generation of lithography tools employs extreme ultraviolet (EUV) light with a wavelength of 13.5 nm to create silicon features of a few nanometres. The demand for more effective clean rooms and monitoring systems is growing as the semiconductor industry transitions to extreme ultraviolet (EUV) lithography, which increases the need to decrease airborne molecular contamination (AMC). [82], [83],[84] . An example of the deterioration and the consequent quality defect is highlighted in Fig.5.8 This initial deteriorating state demands tighter control over humidity and temperature. The lithography process can develop flaws as a result of dust from the air landing on semiconductor wafers and lithographic masks. Semiconductor firms are being forced to decrease the ISO classes of their cleanrooms in order to comply with the increased need to eliminate airborne pollution. To reach the desired particle count, use particle filters and air exchanges. An ISO Class 5–7 cleanroom for semiconductor lithography typically has temperature control set to 20 °C 0.01 °C and humidity set to 45 % 5 %[83]. Even with a temperature variation of 0.1 ºC significant errors may occur. However, by maintaining the temperature variation to ± 0.01 ºC, the precision of a typical lithography operation can improve by an order of magnitude.

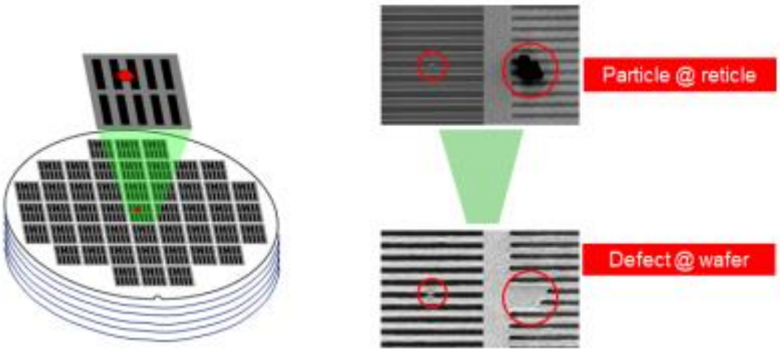


Figure 5.8: Particle dust contamination example

5.9. Lens aberration

The definition of a perfect lens is one that projects the incoming light waves as a single point on the picture plane. The focus point is where attention is drawn. The light waves diverge as a result of aberrations, blurring the image. The difference between a lens with no aberrations and one that has them may be noticed in Figure 5.9. There are several problems in the machine that causes aberrations in the lens, the main problem is lens heating. When the machine is producing wafers, energy from the laser heats the projection lenses. This changes the optical properties of the lenses. The lenses will reach a thermal equilibrium. At that point the optical properties and aberrations remain constant. When the machine stops the lenses cool down and the optical properties change again. To solve these kind of issues the lenses are controlled by a feed-forward and feedback mechanism. When a failure happens the system can compensate for some failure as stated in [7], [79]. If a part is slowly breaking it might be that at the start the systems can correct this defect. After some time the part completely breaks and the system cannot correct it anymore. From the moment it starts to fail the machine is not performing optimal anymore. This should be detected as soon as possible to prevent the machine from malfunctioning.

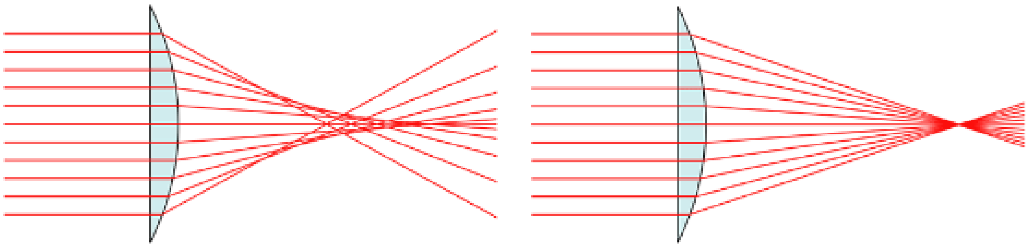


Figure 5.9: Lens aberration on the right side compared to the normal behaviour on the left side

In [79] it is studied the aberration of Wet Exchangeable Last Lens Element (WELLE), which is the last lens in the projection lenses. In this case study SPC and APC is combined. In fact lot of manipulators lens and tools are present in the machine, which are used to control different issues such as lens aberrations. However, these manipulators do not tell us if a part is broken or in which state of aberration has reached the lens. Therefore, to use the manipulators as best as possible the aberrations of the lenses need to be known [79]. This behaviour is described trough coefficients multiplied with Zernike polynomials such that the linear combination describes the aberrations in the lenses. However, the constant manipulator adjustments, as well as the periodic extraction of valuable data and translation of it to lower dimension eliminating extraneous disturbances, provide significant challenges in immediately detecting the true behavior of the process. Therefore, there isn’t a clear choice in how the control actions has be done. [79] proposes a combining SPC and APC methodology for the aberration control.

5.10. Objectives

Methods to find faults in a part are required to deal with all these problems and enhance the machine's diagnostics. Improving machine diagnostics requires automatically determining if a machine is healthy or unhealthy. When a failure's root cause is identified, it may be quickly fixed. Mutually identifying early failure flags depending on the output's quality would be much better. such that it may be resolved without being aware of the predictable evolution of the process failure. These failures could be caused by problems like organic contamination drift, calibration issues or wear. To detect these failures thousands of signals are being measured both from the machines and the multivariate SPC control. This data could be used for automated monitoring to timely estimate the health of the system.

The objective of this thesis is to develop and simulate a methodological framework in which both machine and quality control are integrated to predict unobservable tool degradations under variable operating conditions. This work analyses how the manufacturing system engineering impacts over the level of visibility of the in-situ process monitoring designed and the product quality information. To make this problem clearer we answer the following questions:

- What feedback information has the priority to preventive interventions?
- According to what condition of manufacturing design this priority changes?

To make the goal clearer, following outcome is formulated:

- Define a simulation model that integrates APC and SPC control
- Define the policy control threshold on machine
- Define the policy threshold on the quality control station
- Analyse how the main performances of the line variates with different policy threshold combinations
- Determine the trade-off of implementation of the two policy
- Analyse how this trade-offs changes with a different manufacturing system configuration

6 Simulation model description

6.1. Description of the reference system

The reference system deals with the analysis of a serial manufacturing system with an on-line remotely inspection station. The inspection station, as is common in semiconductor manufacture, is recognized as the system's bottleneck, limiting the rate of production. Upstream machines suffer from an increasing and partially observable degradation of their equipment over time, which has a direct impact on wafer manufacturing quality requirements. However, the machines presented in Fig. 6.1, experiences random events such as failures and repairs. Given the unreliability of the production system, buffer stock is required to defend against backlog during the periods when the system is unavailable due to interruptions. In response to each failure event, a repair intervention can be conducted, which returns the machine to an as-good-as-new conditions.

The proposed quality control policy implies that a sampling fraction of produced items is inspected before being transferred to the inventory stock. Once defective items are identified upon inspection, if the lot inspected overcomes the limit of defected items allowed, the entire lot is thrown away and the next lot is inspected. Depending on the proportion of defectives found in the inspection, the decision maker can decide to immediately initiate one maintenance policy (CBM). On the degrading upstream machine, a predictive maintenance (PdM) activity is performed jointly, estimating the deteriorating behaviour with stratified states of degradation forecasted from observable degradation signal from previous information obtained hypothetically from a sensor placed over the equipment. Such maintenance options enable us to completely mitigate the effects of deterioration on the machine and restore its performance to brand new conditions. The durations of the minimal repair and the preventive maintenance are stochastic and given the set of disturbances that could appear during production, shortages may occur. Anyway, the average preventive maintenance time window is assumed to be shorter that unplanned corrective maintenance. The objective is to determine the production rate, the fraction of production inspected and the optimal joint maintenance policy that minimize the total incurred cost. Total cost includes inventory, backlog, inspection, repair, preventive maintenance, and defectives costs. The optimal solution must ensure that final customers are protected with a constraint on the outgoing quality of items that they receive. A discrete/ continuous simulation model was developed to reproduce the stochastic behaviour of the manufacturing system under analysis. The simulation software Matlab Simulink was used to develop such model. The choice of using a discrete simulation model was used to considerably accelerate the simulation execution time.

6.2. Simulation model

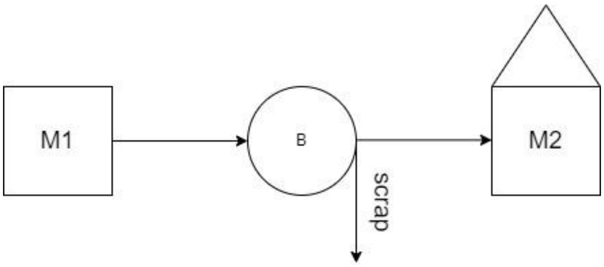


Figure 6.1: Two machine line with positive scrap rate

The model shown in Fig.6.1 considers the last station of the line as inspection M2. It has the role of approving the quality level of operations realized by upstream machines, studying the overall quality parameter requirements that must have the product to be accepted by the final customer. The specification limit is defined as follow:

$$LCS = \mu + 3\sigma \times C_{pk} \qquad (6.1)$$

μ: *average of parameters measured*
σ: standard deviation
C_{pk}: *Capabilty index*

The inspection station identifies a product as faulty if the parameter under examination exceeds the standard limit. Moreover, the inspection station determines also the in-control state of the upstream process, signalling from remote whenever an out of control (OOC) holds through SPC approach. In this model, the SPC control is applied over a generic parameter measurement of the final wafer realized by the upstream machine M1. This measure is assumed to have a normal distribution, whose mean is noised overtime by the upstream machine degradation. As stated in [85] for the positive scrap rate model, whenever an out-of-control (OOC) is detected, a signal is sent remotely from the inspection station to the machining, which is stopped for an external intervention to reset the machine and restore the process to as-good-as new condition. EWMA chart is adopted as R2R approach to highlight in time the process deviations from its starting nominal condition.

The Exponentially Weighted Moving Average (EWMA) is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time. For the Shewhart chart control technique, the decision regarding the state of control of the process at any time, t, depends solely on the most recent measurement from the process and, of course, the degree of "trueness" of the estimates of the control limits from historical data. For the EWMA control technique, the decision depends on the EWMA statistic, which is an exponentially weighted average of all prior data, including the most recent measurement.

By the choice of weighting factor, λ, the EWMA control procedure can be made sensitive to a small or gradual drift in the process, whereas the Shewhart control procedure can only react when the last data point is outside a control limit.

The statistic is calculated with the following equation [86]

$$EWMA_t = \lambda X_t + (1 - \lambda) EWMA_{t-1} \quad for\ t = 1,2, \ldots., n \qquad (6.2)$$

Where:

- $EWMA_0$ is the mean of historical data
- X_t is the observation at time t
- n is the number of observations to be monitored including $EWMA_0$
- $0 < \lambda \leq 1$ is a constant that determines the depth of memory of the EWMA

The parameter λ determines the rate at which "older" data enter into the calculation of the EWMA statistic. A value of λ=1 implies that only the most recent measurement influences the EWMA (degrades to Shewhart chart). Thus, a large value of λ (closer to 1) gives more weight to recent data and less weight to older data; a small value of λ (closer to 0) gives more weight to older data.

The estimated variance of the EWMA statistic is approximately

$$s_{ewma}^{2} = \frac{\lambda}{2-\lambda}\sigma^{2} \qquad (6.3)$$

when t is not small and where s is the standard deviation calculated from the historical data. The center line for the control chart is the target value or EWMA0. The control limits are:

$$LCU = EWMA_0 + ks_{ewma} \qquad (6.4)$$

$$LCL = EWMA_0 + ks_{ewma} \qquad (6.5)$$

where the factor k is either set equal 3 or chosen using experimental tables. For the implementation of this chart, data are assumed to be independent and these tables also assume a normal population[87].
The control chart descripted is shown in the Fig.6.2 below:

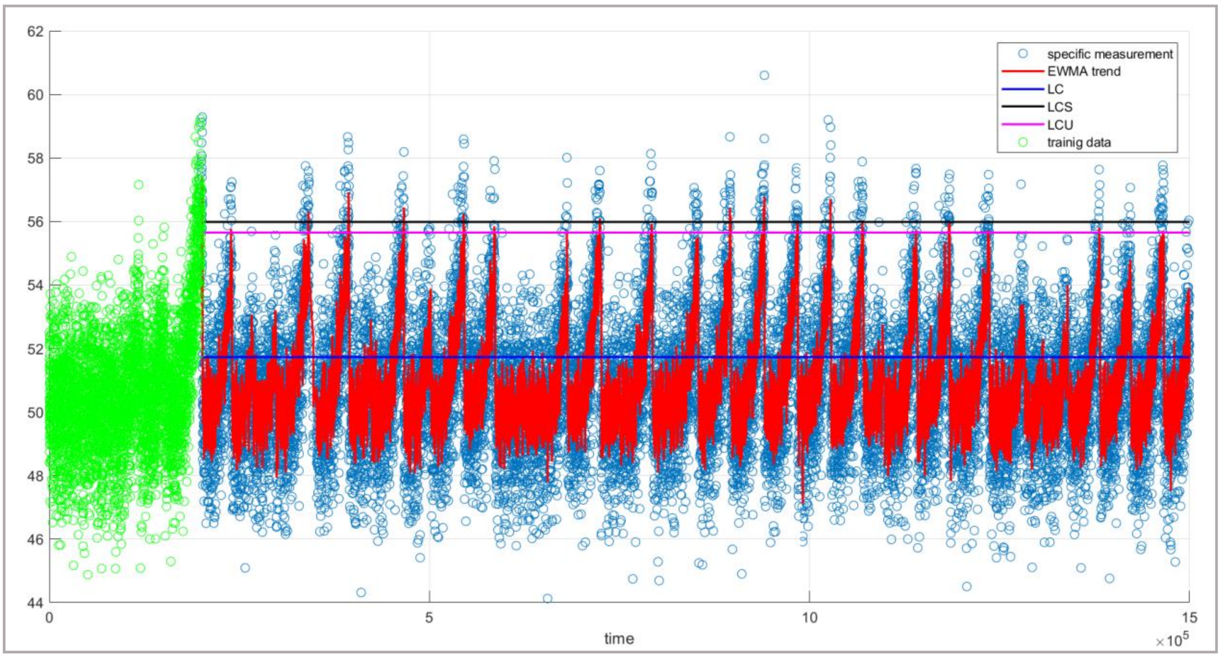


Figure 6.2: EWMA control chart

As highlighted in the Fig.6.2, the EWMA control limit is determined from training data during the warming period of the simulation. The performance of the designed model will be assessed afterwards. Since the increasing degradation state shift upwards the average of the process close to the superior specific limit (LCS) just the upper limit control of EWMA chart is accounted (LCU). Once the performance overabounds LCS the part is considered defective. If above the half of the sample is detected as defective, the entire lot is thrown away. The efficiency of the inspection is assumed to be 100%. Every time the measure is inspected from a piece of a lot the related EWMA is adjourned overtime. once the EWMA trend goes beyond the LCU threshold an OOC is signalled to the upstream machine to reset the process from the origin condition. The inspection policy adopted is the fractional inspection. Instead of making control quality over the entire batch, a random sample of the lot is examined. If the sample detects a limit threshold of defects within a sample the entire lot is scrapped. Intuitively, fractional inspection allows to be faster in the quality procedure as the performance as well. The full inspection, instead, is perfectly accurate to scrap all the flowing scrap product. On the other hand, the time required to implement this policy would reduce the throughput of the entire system.

As previously described, upstream line processes often are subject to continuous deterioration of their equipment that may affect qualitatively the outgoing output result, sometimes with an unknown relation between the process condition health and the quality result.

For the analysis of an integrating predictive maintenance and quality that aims to cope this issue , the upstream machine is characterized by 2 failure mode: One failure exponentially distributed occurs frequently and is repaired with a minimum cost of time, with expected value respectively equal to $MTTF_{M1,failure1}$ and Time to Repair, is exponentially distributed, with expected value equal to $MTTR_{M1,failure1}$. The second failure is longer, it is distributed as a Weibull that provides important estimation of the health of an equipment subject to deterioration. By using Weibull Analysis, there are various indicators, which help in understanding the health of an asset and gives the remaining life estimates. So, in case of scenario where limited data is related to asset life available, we can use Weibull Analysis to understand the Remaining Useful life of an asset. Thus, it makes the Weibull Analysis a good candidate where there is availability of limited data on the asset. Once this last failure mode occurs, the corrective repair would have a huge impact so that why in this model we have to argue on a proper maintenance policy. The hazard function of the Weibull distribution is the input for the degradation behaviour of the machine. This function is approximated with a quadratic function as follow:

$$\text{deg} = K1 * \left(F(t)\right)^2 \qquad (6.6)$$

with K1: given coefficient

F(t): the failure rate overtime

To emphasize the complex, multiple-cause level of degradation a normally distributed random noise factor is introduced. Degradation and failure rate behaviour are presented in Fig.6.3 and Fig6.4.

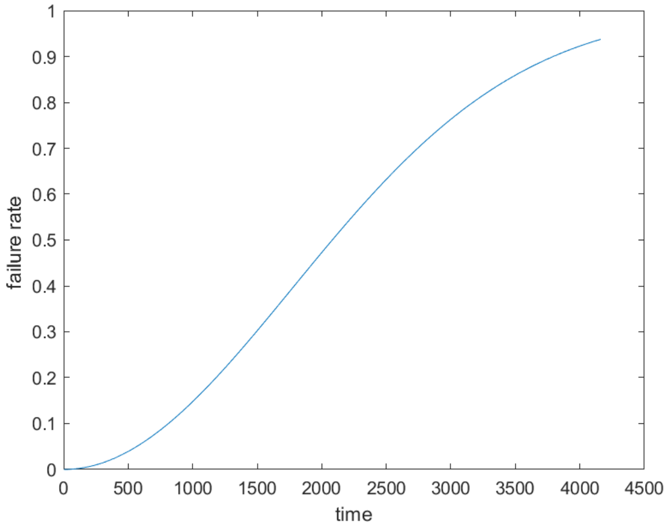


Figure 6.3: Hazard rate

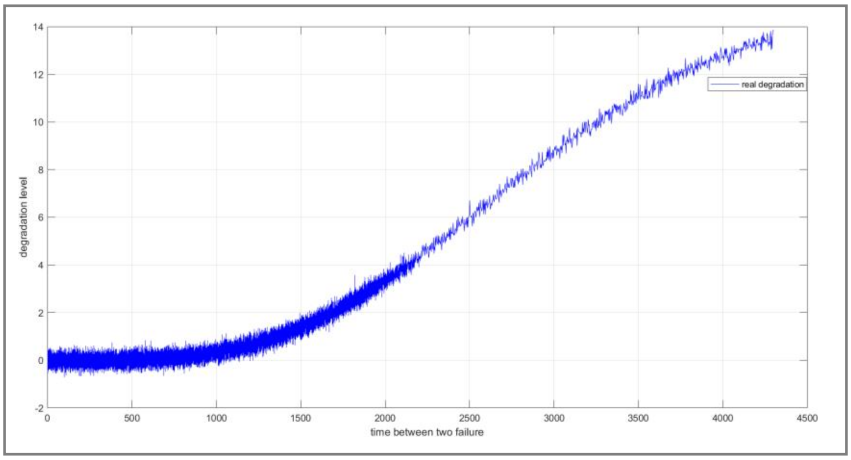


Figure 6.4: degradation level

As the value of degradation level indicator increases, the system performance is getting worse. The exponentially deteriorating trend is employed to imply a physical system that tends to degrade faster as its condition becomes worse. The degradation level is assumed to be partially observable from local sensors. Moreover, its real level will affect the average normal distribution of the outgoing quality parameter over time. the hidden relation is described below:

$$X \sim N(\mu + deg; \sigma^2) \qquad (6.7)$$

X=quality parameter

$\mu = average\ of\ the\ process\ in\ a\ normal\ condition\ of\ the\ equipement$

$\sigma = standard\ deviation\ of\ the\ process\ in\ a\ normal\ condition$

The system presents some other characteristics:

- Once repaired from the mode in which it failed, each machine is restored to as-good-as new condition.
- The line is asynchronous, machines can start or finish a part at any time instant without synchronization with other stations since *M-1* buffers decouple the production pace of the stations
- Each machine has its own deterministic service time
- Only one part is produced at a time
- The buffer capacities are finite
- Blocking After Service (BAS) is defined
- FIFO is adopted

In addition, other assumptions to the following model are formulated:

<u>General assumption</u>

- The presence of defective parts entering the first machine is not considered
- The material stream is unique
- The upstream machine is never starved, and downstream machine is never blocked

<u>Other assumption on the machine M1</u>

- machine can have different operational states, with different service rate
- Service times include the time to load unload the part
- The machine can have several failure modes
- A machine can fail in only one mode at a time
- The processing rate of the inspection machines is calculated dividing the time to measure the m sampled parts between all the *h+m* products

The simulation model has been designed in *SimEvents,* the Discrete Event simulation tool of *Simulink,* which is the dynamic system simulation package integrated in MATLAB.

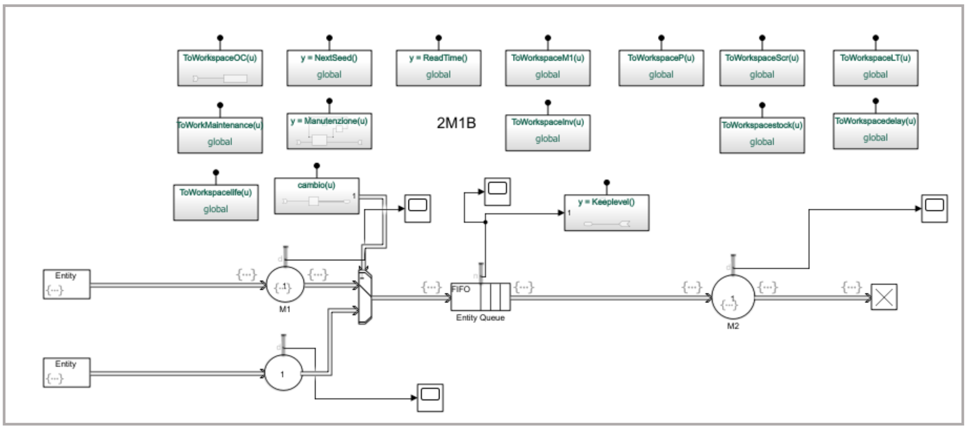


Figure 6.5: Simulink Sim Events model

The model is composed by two machines, one machining station and one inspection station that also perform the scrap of parts.
To analyse how the integration of the new policy can improve the performance of the line, some characteristics are introduced inside the model. Starting from the simple model 2M1B, the upstream machine will have two operation dependent failures. The occurrence of a failure mode-i of a machine-m is accounted by the Remaining Useful Life variable $RUL_{m,i}$. Every time the machine process one part, the $RUL_{i,m}$ is decreased by one service time. As soon as the RUL is smaller than the service time for the next part, the failure mode-i occurs. All $RUL_{m,i}$ are randomly sampled. Respectively, the first failure is sampled with an exponential distribution with mean equal to $MTTF_{M1,failure\ 1}$. The second, instead with a Weibull distribution with scale factor equal to $MTTF_{M1,failure\ 2}$ and shape factor α=2.

Given that the manufacturing system is subject to deterioration, our model seeks to identify the impact of such deterioration process on product quality and read the effects of quality-deterioration on the EWMA control chart.

Together with the quality control policy of EWMA another policy is implemented directly on the upstream machine. As already mentioned in this section not all the physics of the processes are known deterministically as in semiconductor industries, both for the complexity of the sequence of operations and the stochasticity of the event. This is why sometimes sensor devices are needed to have a rough estimate in the equipment condition.
The model designed proposed a stratified level of degradation to represent the state of the machine condition, rather than trying to postulate the exact value of degradation level indicator. One way to discretize the continuous degradation process is evenly divide the vertical range into N regions and each of them corresponds to one degradation state, where N is the number of states used to describe the entire degradation process. Figure 6.6 illustrates the idea of using 5 stratified states to represent the stochastic degradation process shown in Figure 6.6, in which the green line represents the discretized state. It can be seen that as the degradation process evolves, the state number changes from 1 to 5. Furthermore, as we discussed earlier, the exponentially deteriorating trend mimics the fact that a physical system tends to degrade faster as its condition becomes worse. Therefore, it can be seen that the duration that the system stays in a preceding state is longer than that it stays in a succeeding state.

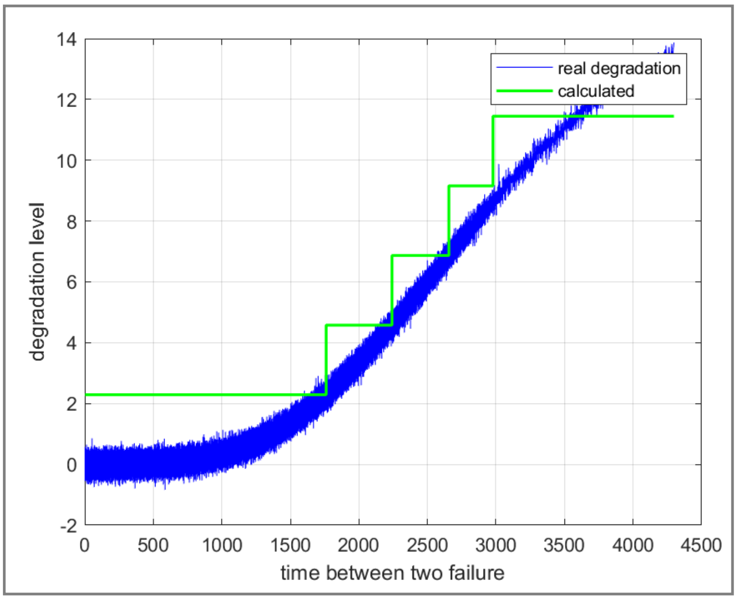


Figure 6.6: Hidden Markov model

Ideally, if one can observe the degradation process shown in Figure 6.5 and further discretize it into stratified states, a PdM decision could be easily made according to the current degradation state. Unfortunately, in most of cases the degradation process is not directly observable, such as the situation in a chamber tool we mentioned in the previous section. Therefore, one will have to rely on the readily observable signals emitted from the deteriorating system to infer the underlying degradation process. In this simulation model, the observable signals are generated as follows. We assume that there is one observable variable emitting from the system. This assumption will only make the simulation model simple, but the HMM modelling procedure will remain the same if there are more observable variables, provided that an adjusted BaumWelch algorithm will be used in multi-sensor cases [131, 132]. We denote '1' to represent 'conforming' signal and '2' to represent 'nonconforming' signal. In reality, the observable signals will be continuous variables in most cases, such as temperature, pressure, and gas flow, which need to be transformed into discrete emission symbols by discretization. As we assume that in the early states the system is in 'good' condition, it tends to generate more 'conforming' signals rather than 'nonconforming' ones; and as the system degrades, more 'nonconforming' signals will occur. An emission probability matrix is provided to generate these types of signals overtime. An example of emission probability table is shown in Table 6.1, in which each row corresponds to a system degradation state, and each column corresponds to the probability of generating one type of emission symbols. For instance, when the system is in state # 1 (best state), it has 90% probability to generate 'conforming' signals denoted by '1', and 10% probability to generate 'nonconforming' signals denoted by '2'; however, when the system deteriorates to the fifth state (worst state), it only has 10% probability to generate 'conforming' signals and 90% probability to generate 'nonconforming' signals.

Table 6.1: Emission probability

	P(observable)	P(not observable)
state1	0.9	0.1
state2	0.7	0.3
state3	0.5	0.5
state4	0.3	0.7
State5	0.1	0.9

Once we have established the underlying degradation states and the emission probability table, a series of observable emission symbols can be generated using the simulation model, which will be used to train a Hidden Markov model (HMM). This model is proposed to overcome the need for direct observations of degradation of the machine, postulating the assumed the deteriorating level progression based on available process information. The HMM is chosen because it is a natural extension of observable Markov chains in which states of the Markov chain are not directly observable and can only be inferred through another stochastic process that describes the sequence of observed states.

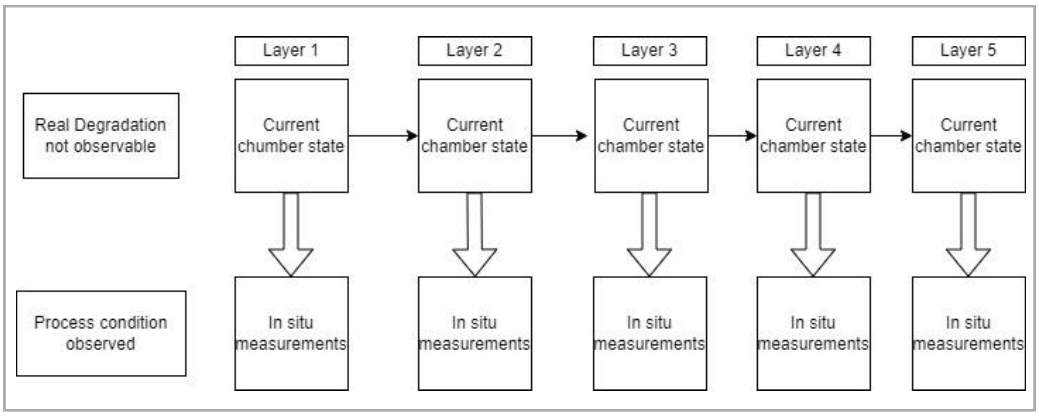


Figure 6.7: Hidden Markov model

In the proposed HMM modelling approach as depicted in Figure 6.7, the directly unobservable state of the equipment will be modelled using observable controllers, in-situ measurement variables, such as temperature, pressure, gas flow, energy consumption.

In Fig.6.8 the topology of the hidden Markov model is proposed to estimate the stratified level of degradation from observable signals

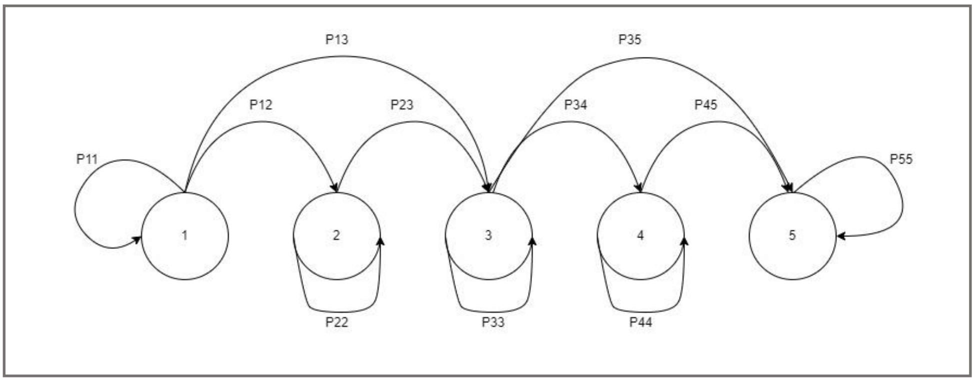


Figure 6.8: Illustration of 5 state hidden Markov chain

Each circle represents a degradation state. Edges along with arrows represent the directions of state transitions, and then the likelihood of this transition happens is depicted along with each edge. For instance, P11 means the probability that state # 1 will stay at its current state; P12 means the probability that state # 1 will transit to state # 2, P13 means the probability that state # 1 will transit to state # 3, and so on. The transition probability matrix and emission are assumed give from previous experimental campaign.

$$transition\ probability\ A = \begin{bmatrix} 0.9975 & 0.0025 & 0 & 0 & 0 \\ 0 & 0.9956 & 0.0044 & 0 & 0 \\ 0 & 0 & 0.9935 & 0.0065 & 0 \\ 0 & 0 & 0 & 0.9920 & 0.0080 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.8}$$

$$emission\ probability\ B = \begin{bmatrix} 0.8888 & 0.1112 \\ 0.6917 & 0.3083 \\ 0.4788 & 0.5212 \\ 0.2801 & 0.7199 \\ 0.0730 & 0.9270 \end{bmatrix} \tag{6.9}$$

The trained HMM can be used along with the observable signal to estimate the underlying state transition, which is plotted against the original degradation indicator as well as the stratified states, as shown in Figure 6.9, in which the solid line represents the estimated states. It can be seen that the estimated states follow the same pattern of stratified states very well except the last state due to the deficiency of number of states.

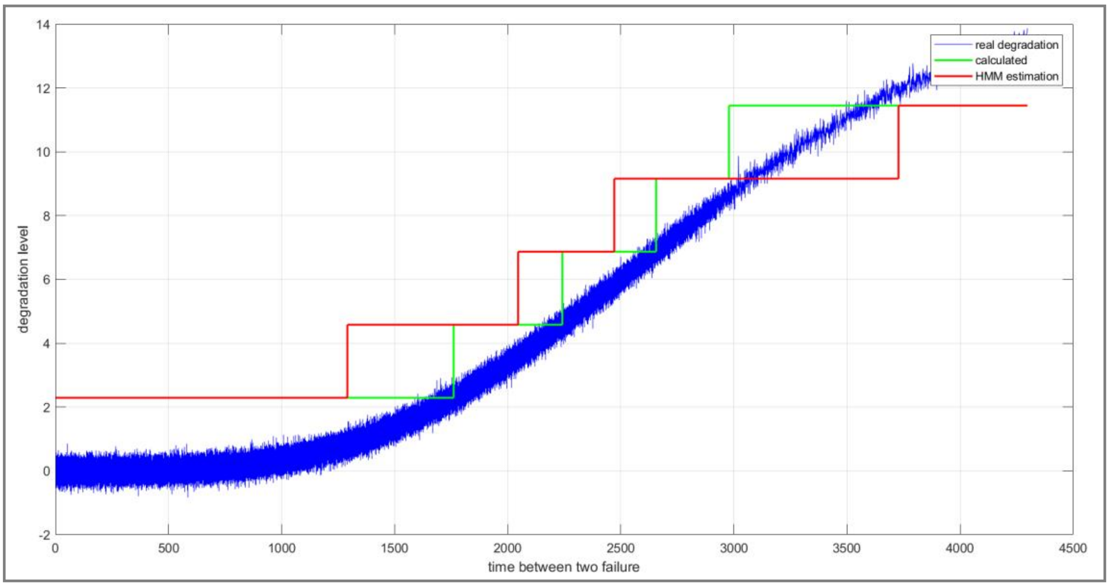


Figure 6.9: Illustration of 5 state hidden Markov chain

In order to evaluate the performance of this model quantitatively, the sum of squared error (SSE) of using stratified state to represent original degradation indicator is calculated as a benchmark reference value.

$$SSE_{stratified} = \sqrt{\sum_{t=1}^{T}(D_t - S_t)^2} \qquad (6.10)$$

$$SSE_{estimattion} = \sqrt{\sum_{t=1}^{T}(D_t - E_t)^2} \qquad (6.11)$$

t= discrete time
$T = time\ horizon;$
D =Real degradation level;
St= value of stratified function
Et= value of estimated stratified function

Then we calculate the error introduced in this modelling approach by Equation (6.12), imposing a limit of error allowed about 30%.

$$Modelling\ error = \frac{|SSE_{stratified} - SSE_{estimation}|}{SSE_{stratified}} \times 100\% \qquad (6.12)$$

During the model learning process, the state transition matrices A and emission probability matrices B for each HMMj need to be calculated using a training dataset, consisting of sequences of observable variables, such as temperature, pressure, gas flow, and energy consumption. In this model, these two structures are assumed as a given parameter from previous learning process. Then the trained HMM along with observable signals can be used to estimate the states transition of the underlying degradation process. Finally, the estimation results and actual degradation process will be compared to verify the modelling accuracy. This methodology usually is applied a multi-dimensional HMM that has more than one observation symbol at each time [131, 132]. Since it can accommodate different sensor signals simultaneously and transfer all informations contained in the sensors into model parameters, the multi-dimensional HMM is preferably used in this research to fuse multi-sensor data together resulting in better estimation of the tool degradation.

However, the original Baum-Welch or Viterbi algorithm from which matrix A is estimated, is designed for one-dimensional HMM, which only has one observable variable, and therefore needs to be modified to accommodate multi-dimensional sensor data. There are in literature several approaches to deal with this problem[19], [88], [89] but, for sick of simplicity the overall observable degradation is assumed to be determined from one single variable.

Another strong assumption is the fact that machine can process just one type of wafer recipe (single operating condition). With this assumption the degradation function of the machine won't be affected by the complexity of multiple recipes for the moment, since it is not the objective of this study.

In speech recognition [90] and machine condition monitoring applications, 3-state HMMs are often used, which generally yield results that are good enough to represent the corresponding processes. However, since our modelling purpose is to facilitate maintenance decision-making, 3-state HMM does not give enough representation to an entire degradation process. For example, if state # 1 denotes the initial state of chamber performance right after maintenance and state # 3 denotes the state of chamber performance, which is no longer qualified to produce any products. Then the only choice for conducting maintenance is state # 2, which will lead to a trivial solution of maintenance decision-making. Therefore, in order to accommodate the maintenance decision-making representation, the degradation process using 5-state HMM is designed in order to select the one that yields the maximum likelihood estimates.

It Is used a MATLAB script to set the line characteristics (machines and buffers), and also to save the result.

The configuration parameters are:

1. Buffer capacity : N
2. Service rate the upstream machine μ_u
3. Service rate of downstream machine μ_d
4. Failure rates upstream machine p_{1u} , p_{2u}
5. Repair rates of upstream machine r_{1u} , r_{2u}
6. Number of inspected parts m
7. Lot size l
8. Cost for unplanned intervention Cc
9. Cost for preventive maintenance Cp
10. Cost for a single scrapped part Csc
11. Cost for a single inspected part Cinsp
12. Cost for OOC intervention Cooc

The fixed parameters for this case are reported in the Table 6.2. In this paradigm, the upstream production rate mimics the lithographic production rate. According to [7], wafers are typically made in lots of 25 pieces, and since a photolithography machine produces one lot every 66 minutes, it takes 2.64 minutes to make a single wafer. Whether under control or not, this machine's production rate is the same for all of the up-states; as the production rate is determined by the quantity of pieces created in a given amount of time, its value is:

$$\mu_u = \frac{1}{2.64} = 0.3787 \ \frac{parts}{min} \quad (6.13)$$

If all the 250 candidate points are measured, the inspection time for a single wafer is 1 hour. For each lot, it is inspected a number m of wafers, so the total inspection time for each batch is m hours. This total time is equally divided to all the wafers in the lot, so the inspection rate becomes[7]:

$$\mu_d = \frac{1}{t_{inspection} \times (\frac{m}{h+m})} = 0.01667 \times (\frac{lot\ size}{m}) \ \frac{parts}{min} \ (6.14)$$

Table 6.2: system parameters

μ_u	lot	p_{1u}	p_{2u}	r_{1u}	r_{2u}	p_2	r_2
0.3787	25	0.008	0.0004	0.05	0.002	0.01	0.1

The simulation settings are reported in the Table 6.3 below:

Table 6.3: simulation setting

	value
Runs	10
Run length	1000000 t.u.
Warm-up length	200000 t.u.
Confidence level	0.025

Preliminary runs are performed before every simulation settings to build the EWMA chart for the inspection policy and stratified degradation levels for the preventive maintenance over the machine.

6.3. Simulation output

The main performances assessed from the system are the following:

- Throughput effective: $THeff = \frac{good\ parts\ produced}{time\ horizon}$
- Throughput total: $THtot = \frac{total\ production}{time\ horizon}$
- Yield rate: $Yield = \frac{THeff}{THtot}$
- Delay feedback: time interval between the occurrence of OOC and the alarm of it from the inspection station
- Number of corrective interventions (Nc): frequency of unplanned intervention
- Number of preventive interventions (Np): frequency of intervention activated respectively by machine control and quality control policy
- Number of OOC (Nooc)
- Expected total cost: $ETC = E(TC) = \frac{TC}{time\ horizon}$
- Total cost: $TC = Nc \times Cc + Nooc \times Cooc + Np \times Cp + Nscrap \times Csc$

6.4. Summary

In MATLAB, the model is represented. Simulink creates a 2M1B line in which the downstream machine serves as the serial line's inspection quality and bottleneck. The machining operation is performed by the upstream machine, which is susceptible to two stochastic failure modes. With a Weibull distribution, one failure in particular deteriorates over time. The linked degradation has an effect on product quality that is not visible to the system. The upstream machine's deterioration is detected by indirect and discretized signals, which are transformed into a stratified function using HMM, which assesses the level of degradation attained to assess the machine's health. Indeed, the number of stages of degradation after which to intervene determines the machine controller policy. On the control quality point of view, a threshold limit on the exponential weighted average (EWMA) recorded run by run is applied. Once the function overcome the deviation limit a preventive intervention is applied.

The purpose of the model is to investigate the behavior of Maintenance policies mandated by machine controllers and the CBM threshold applied to the EWMA control chart before detecting an OOC.

The model's proposed policy will act as follows: When one of the two controllers triggers an alarm, the system is restored to its original state, resetting both the product's health and quality.

7 Design of Experiment

Design of experiments (DOE) is a systematic, rigorous approach to engineering problem-solving that applies principles and techniques at the data collection stage so as to ensure the generation of valid, defensible, and supportable engineering conclusions. In addition, all of this is carried out under the constraint of a minimal expenditure of engineering runs, time, and money.

There are four general engineering problem areas in which DOE may be applied[91]:

- Comparative: The engineer wants to determine whether a change in one aspect has changed or improved the process as a whole.
- Screening Characterization: In this case, the engineer is interested in "understanding" the process as a whole in the sense that he/she wishes (after design and analysis) to have in hand a ranked list of important through unimportant factors (most important to least important) that affect the process.[92]
- The engineer wants to "understand" the process as a whole, therefore after design and analysis, he or she hopes to have a prioritized list of elements that have an impact on the process (from most important to least essential).

Among the most prominently used DOE techniques are Response Surface Methodology with Central Composite Design, Taguchi’s method and Factorial Design.

Factorial design is used for conducting experiments as it allows study of interactions between factors. Many processes are driven by interactions. Without a factorial experimental design, an important interface can be unnoticed. In a complete factorial experiment, responses are assessed at all combinations of the experimental factor levels. The conditions under which responses are measured are represented by the combinations of factor levels. While factorial design may be done on two-levels, three-levels, and multi-level factorial, each experimental condition is referred to as a "run," and the response measurement is referred to as an "observation." The "design" is the full collection of runs. Full Factorial Design is a design in which all potential combinations of the factor levels are satisfied. Full factorial tests would produce more trustworthy results, but they are expensive and occasionally impractical to undertake. [93]

Central composite designs (CCDs) [93], also known as Box-Wilson designs, are appropriate for calibrating the full quadratic model described in Response Surface Models. There are three types of CCDs, namely, circumscribed, inscribed and faced. The geometry of CCDs is shown in Fig.7.1

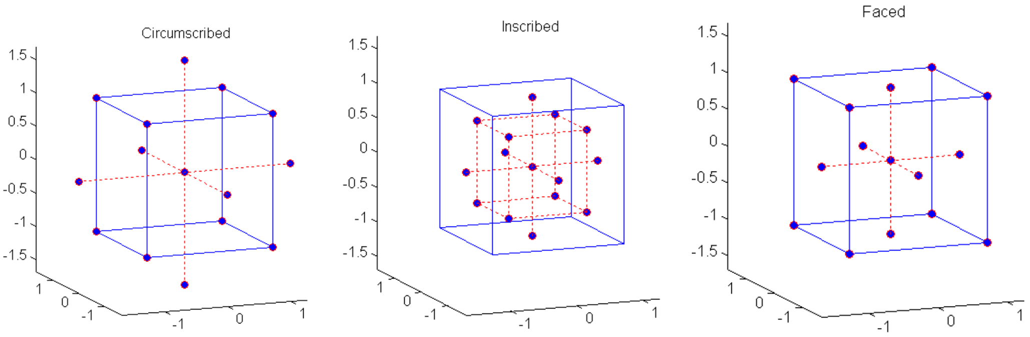


Figure 7.1: Total throughput reduction with sampling rate increase

Each design consists of a factorial design (the corners of a cube) together with centre and star points that allow estimation of second order effects. For a full quadratic model with n factors, CCDs have enough design points to estimate the (n+2) * (n+1)/2 coefficients in a full quadratic model with n factors. The type of CCD used (the position of the factorial and star points) is determined by the number of factors and by the desired properties of the design.

Full factorial and inscribed DOE are implemented for the extraction of significative data in order to get Response Surfaces related to Yield rate, the Throughput effective seen by the customer and Expected total cost.

Simulation runs are conducted according to a complete 3^2(3 factor, 2 levels) factorial design jointly with a CCD inscribed design (16 experiment) to screen out a subset of the control factors: The state of degradation estimated in which intervene and the quality control limit imposed to the EWMA function deviation, which have a significant impact on the ETC(.), THeff(i), Yield(i). For each combination of independent factors the experimental design is replicated three times implying a total of $(3^2 + 16)$*3=75 simulation runs for each response surface. We are interested to take significant factors and build a metamodel of how the simulation model transforms a particular set of input-factor values into the output response. Based on off-line simulation runs we define the minimum and maximum values of the factors as presented in Table 7.1. The simulation results are handled again in Matlab in order to obtain an analysis of variance (ANOVA). Additionally, a residual assessment is carried out to judge the surface fitting and choose a metamodel for the varied performance that is intended to be analysed. These steps are performed for every surface response obtained. An example of the analysis performed are presented in Table 7.2 and Fig.7.3-7.4.

To evaluate the goodness of the approximated response. Generally, The adjusted R-squared coefficient of the data of Table 7.2 for every surface extracted for every configuration falls in a interval between R_{ad}= [92.1%, 96%], with the exception of ETC functions that oscillate between 87.3% and 93%. From Table 7.2 it is possible to note that all the main factor and most of the interactions are significant with a Pvalue≤ 5%. In Fig.7.2 an example of approximated yield in function of the two-threshold level is provided.

Table 7.1:Range for independent variables

Factor	**Low level**	**High level**	**Description**
Machine state treshold	1	5	State of degradation estimated with HMM
EWMA treshold	μ	LCU	Treshold appliead on EWMA function

Table 7.2: Anova analysis

Coefficients	Estimate	SE	tStat	pValue
γ_0	-7.229	1.9946	-3.6243	0.0005498
γ_1	0.2904	0.072637	3.9979	0.00015801
γ_2	0.31417	0.024177	12.995	3.2335×10^{-20}
γ_3	-0.0055383	0.00041648	-13.298	1.0308×10^{-20}
γ_4	-0.0025596	0.00066076	-3.8737	0.00024097
γ_5	-0.0031578	0.00083336	-3.7892	0.00031977

Number of observations: 75, Error degrees of freedom: 69
Root Mean Squared Error: 0.0105
R-squared: 0.937, Adjusted R-Squared: 0.933
F-statistic vs. constant model: 206, p-value = 5.18e-40

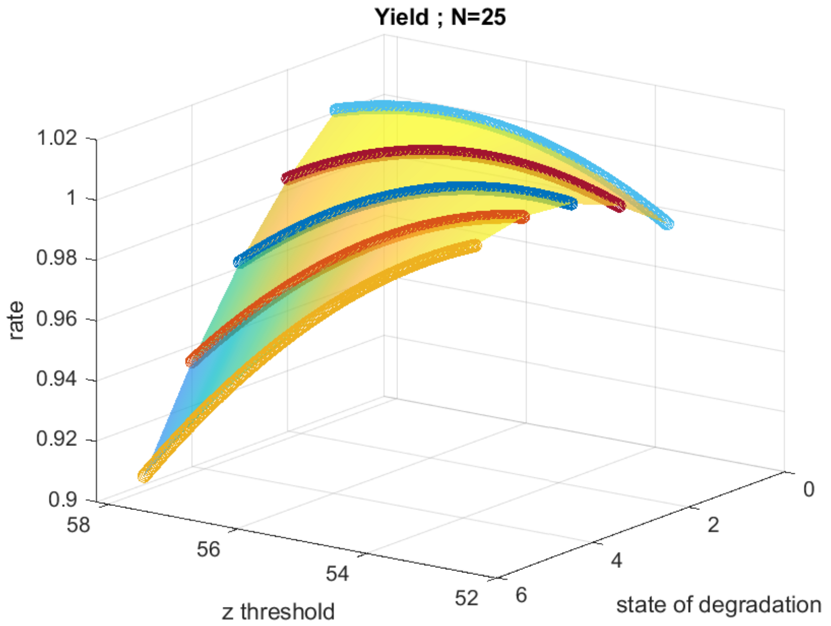


Figure 7.2: Response Surface of yield

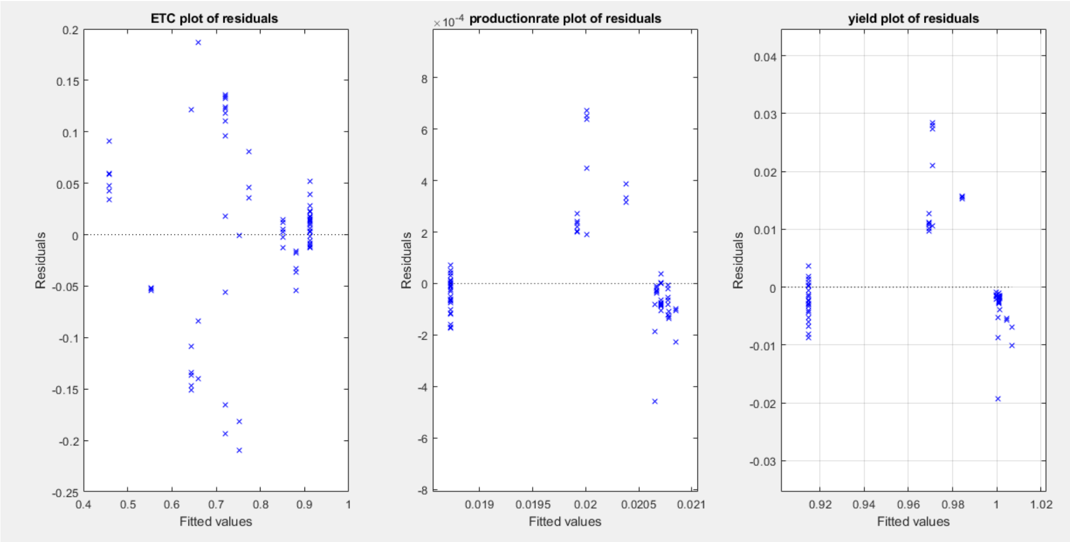


Figure 7.3 Residual dispersion analysis

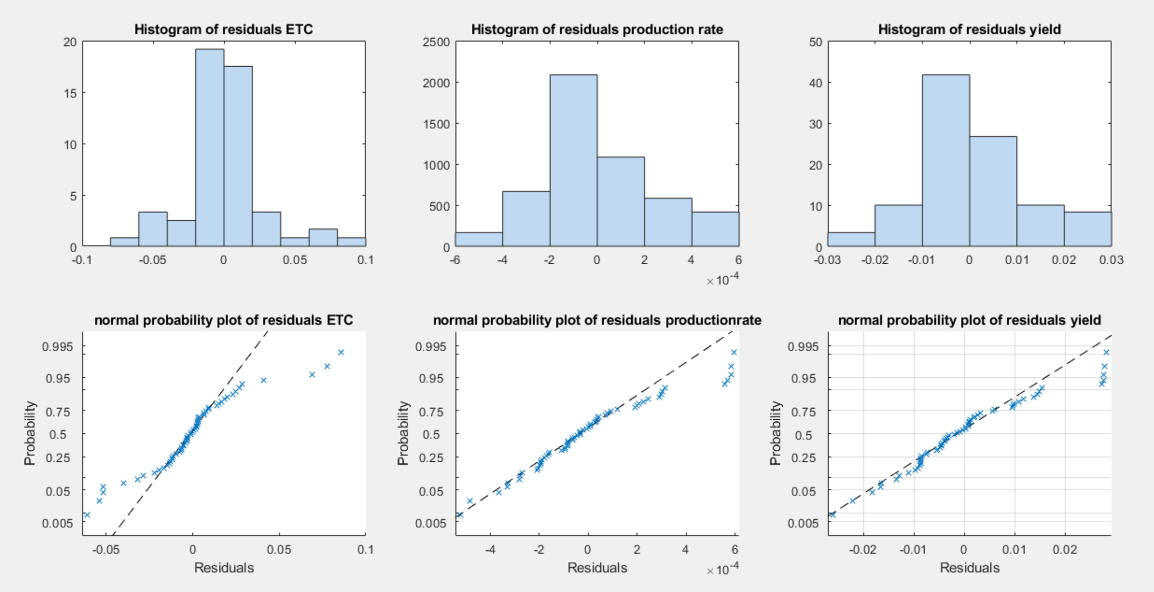


Figure 7.4 Normal distribution check of residuals

8 Numerical results

In this chapter, results from analysis conducted on the model are exposed. The conducted investigation begins with a sensitivity analysis connected to the several major configurations of the line. The trade-off between the machine control and quality control thresholds was then examined in relation to variations in a single configuration. The research comes to a close with the results of several policy optimizations intended to reduce total expenses, maximize yield, and maximize profit.

8.1. Sensitivity analysis

In this phase, the line's sensitivity analysis will be evaluated in order to determine how changes in the key configuration factors influence the line's performance. In order to choose the best integrated quality and maintenance strategy, it is essential to know the system setup in advance.

8.1.1. Sample size variation analysis

The proportion of inspections in a lot is the first configuration to influence policy decisions. In particular, as can be shown in fig.8.1, the system's total productivity falls as the sampling proportion rises.
However, an higher fraction of inspection entails higher knowledge about the process that allows to be more reactive to anomalies. indeed, the first trade off of this first system configuration is about productivity against quality. Moreover, another element that may affect the yield of the system is tied to the delay of signal. In fact, the higher inspection lot increase the WIP along the line that will cause a huge delay before get the feedback of degradation of process towards the machine. Anyway the higher sampling rate will benefit the EWMA chart responsiveness that recognize on time the future deviation of the process reducing the number of OOC occurred. But for higher capacity of the inter operational buffer configuration, this behaviour may lead to worse THeff performances compared to lower size of the buffer.

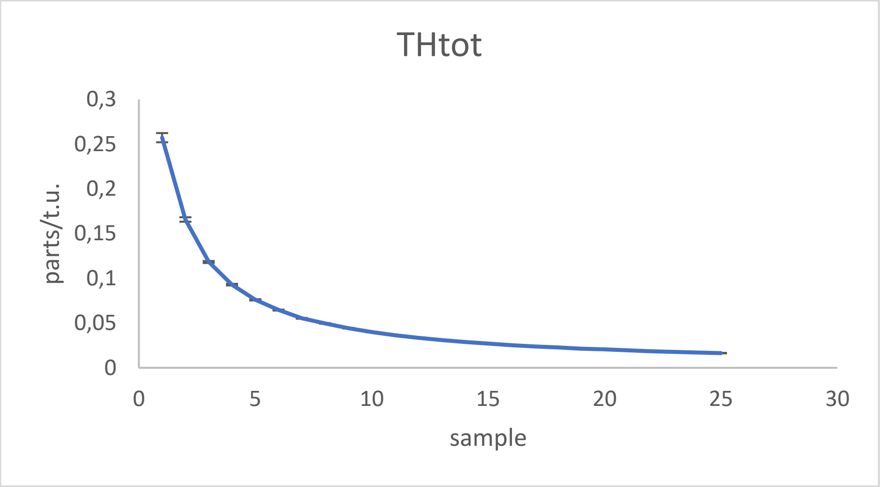


Figure 8.1: Total throughput reduction with sampling rate increase

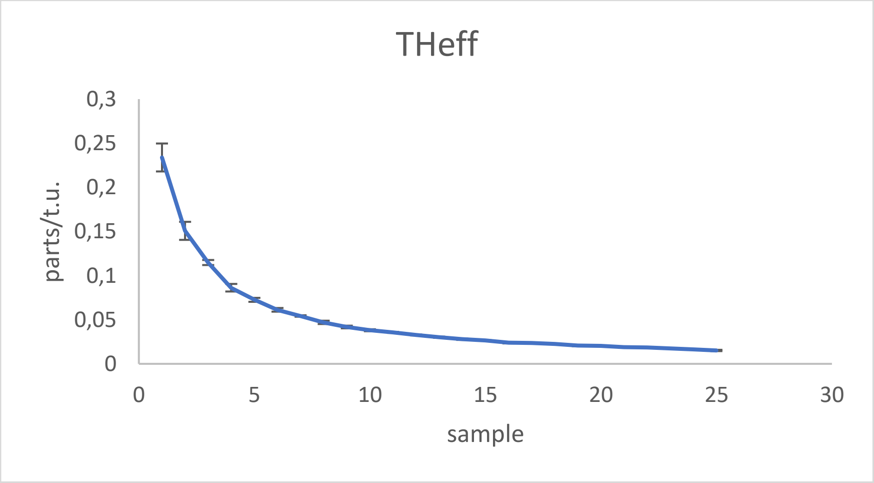


Figure 8.2: Throughput effective

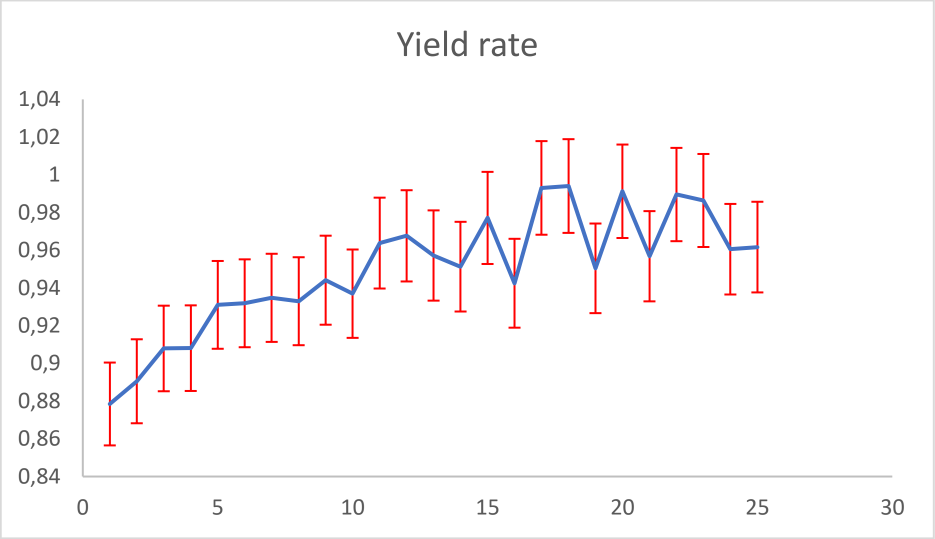


Figure 8.3: Yield rate increase with increasing sampling rate

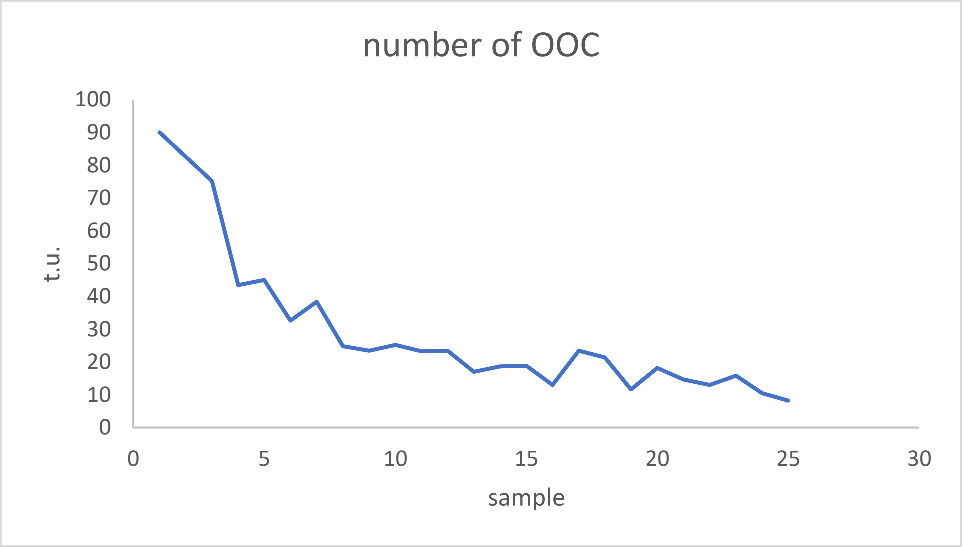


Figure 8.4: Number of Out of control detected by EWMA chart

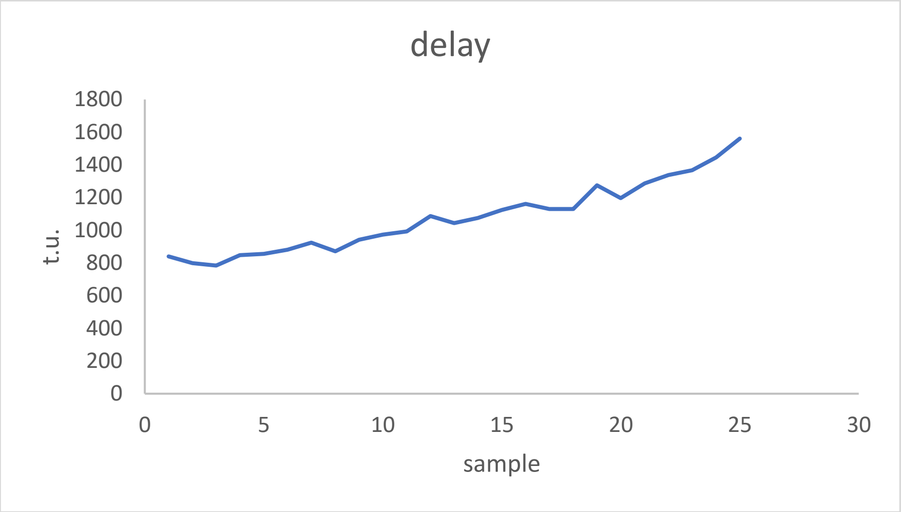


Figure 8.5: Delay of quality feedback

8.1.2. Buffer capacity variation

In contrast to the previous assessment, the random sample fraction strategy used in the simulation has a significant impact on the sensitivity analysis performed over the design of the production system's buffer capacity size. However, it is feasible to partially deduce relevant patterns of the system from this enormous variability that decreases by increasing the sampling percentage examined.

The buffer capacity has a significant impact on the production system's yield rate, as shown in Fig. 8.7 for various sample configurations. This occurs due to two major causes. Because the inspection station is the system's bottleneck, as we already mentioned, the buffer capacity significantly lengthens the signal delay. The number of components stuck inside WIP that need to be examined will increase as buffer size increases. As a result, if any lots that are now in the buffer have measurement discrepancies brought on by the degradation process, they will be detected minutes after the machine degrades. Without any proactive measures to stop this problem, the machine can enter another level of degeneration in the meantime. As a result, increasing buffer capacity results in a significant loss in yield. As can be observed in Fig. 8.13, there are several reasons why performance degrades more quickly with a low sample rate. The upstream equipment might deteriorate more quickly because of the low sample rate even if it has been in use for a long period. Low sampling rates are also less sensitive to significant product quality property deviations.

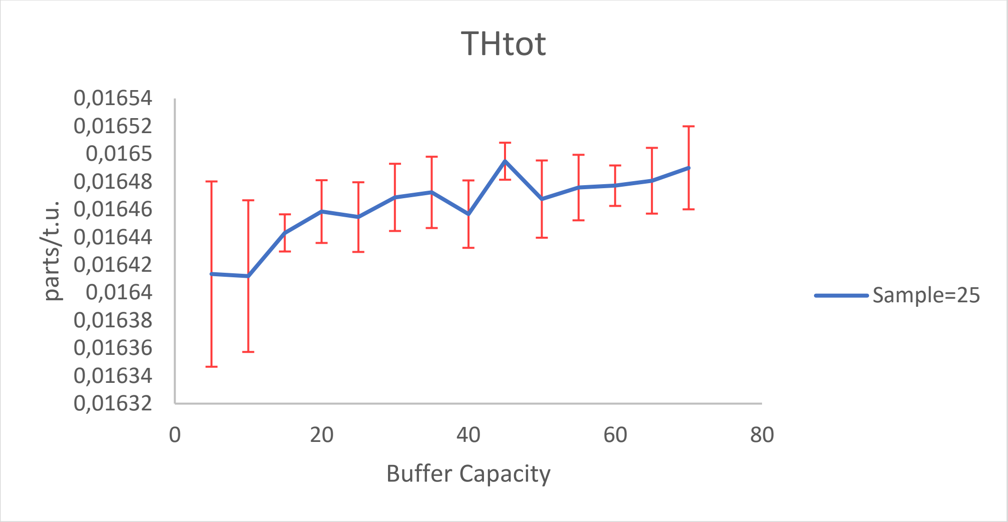


Figure 8.6: Throughput total varying buffer capacity

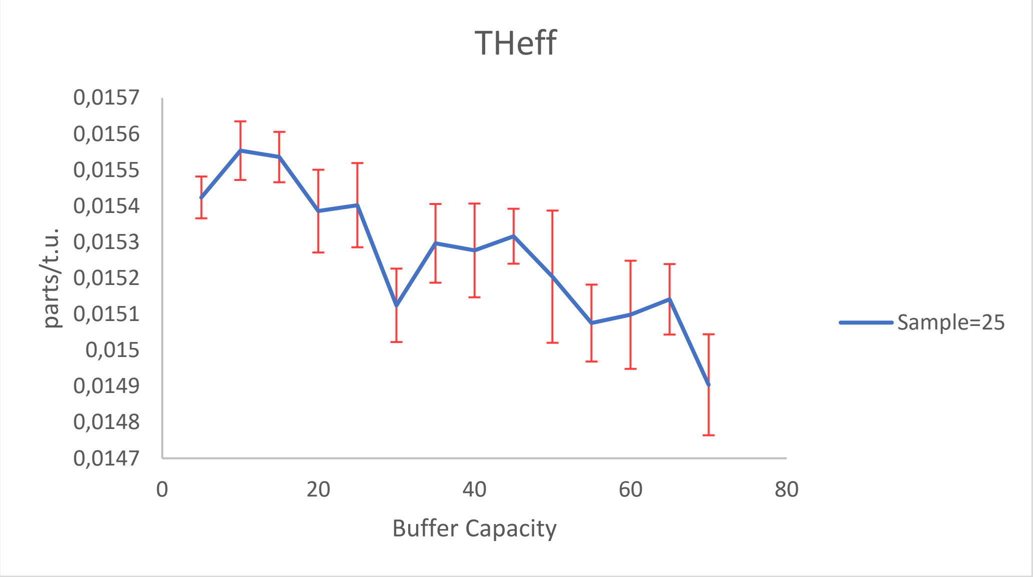


Figure 8.7: Throughput effective

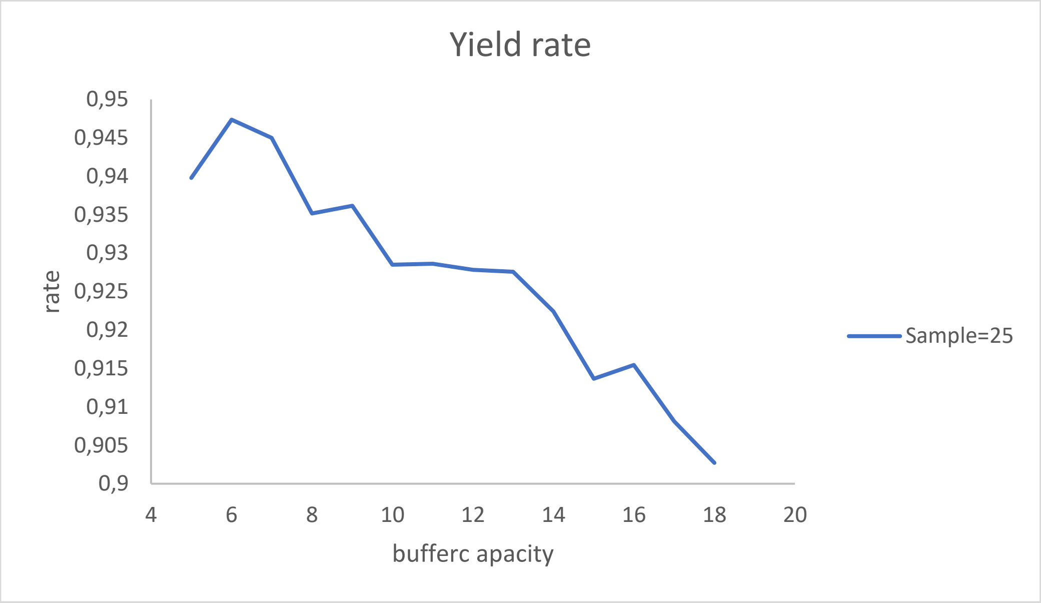


Figure 8.8: Yield rate varying buffer capacity

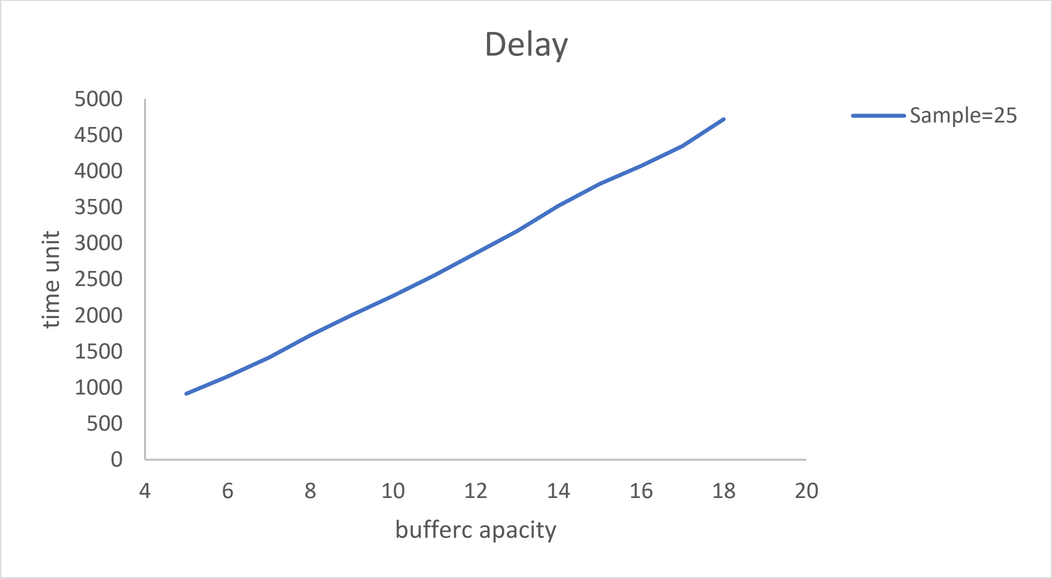


Figure 8.9 : Delay quality feedback by varying buffer capacity

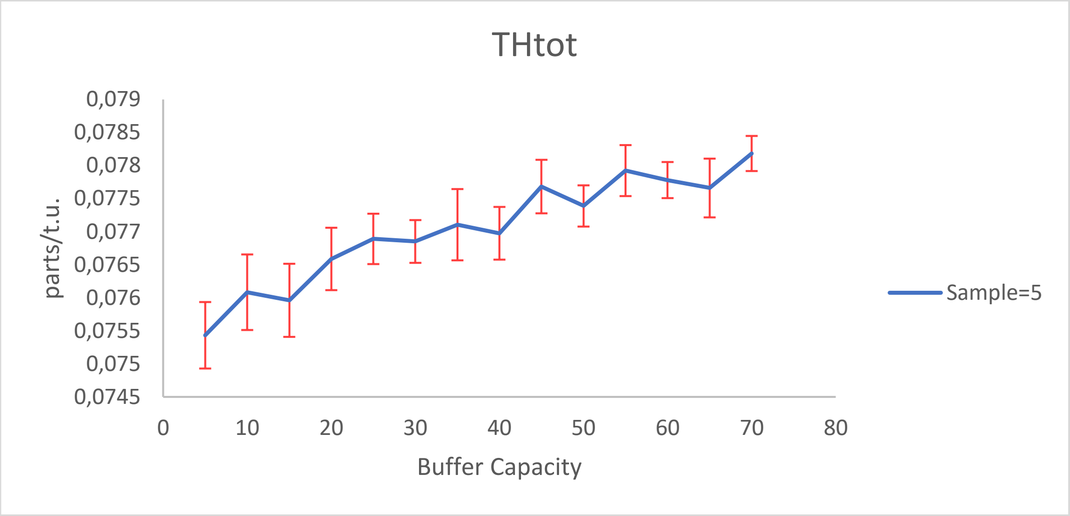


Figure 8.10: Throughput total varying buffer capacity

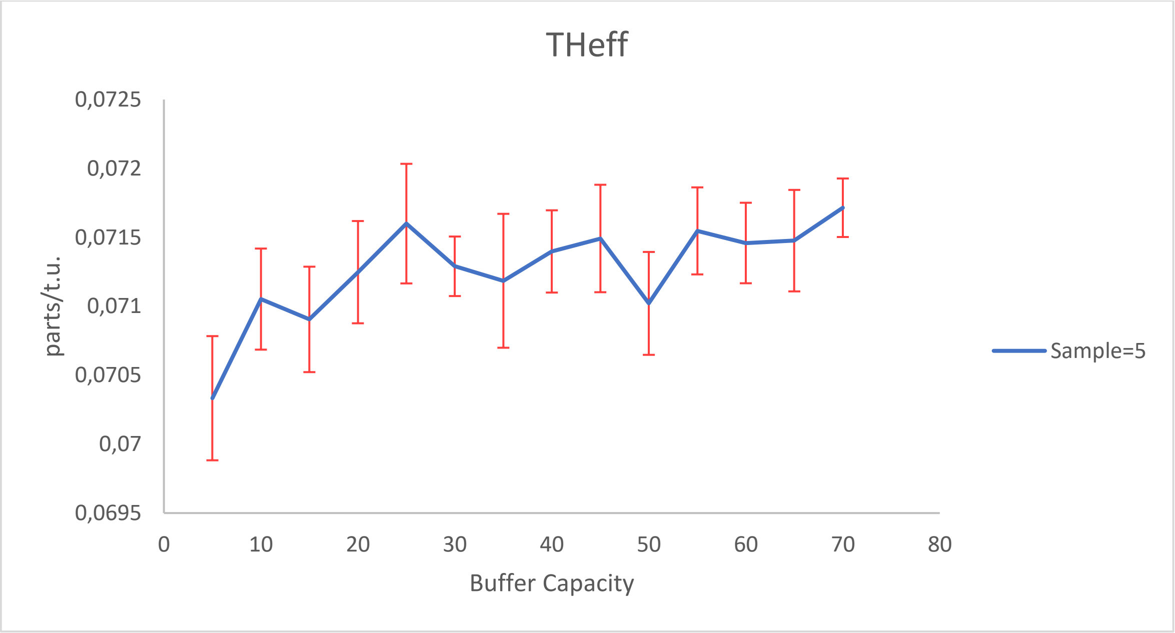


Figure 8.11: Throughput effective

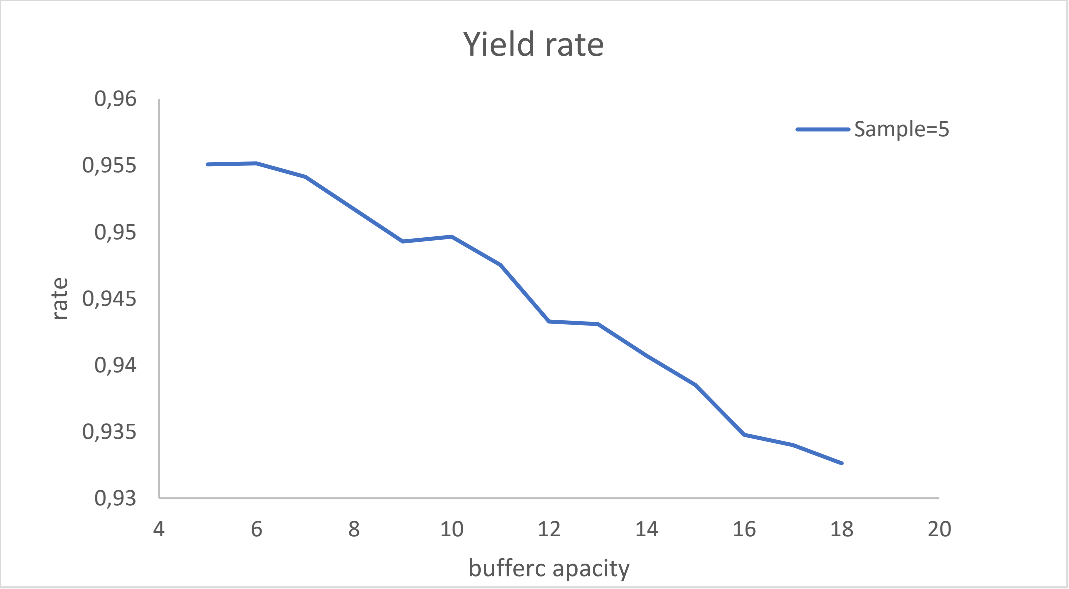


Figure 8.12: Yield rate varying buffer capacity

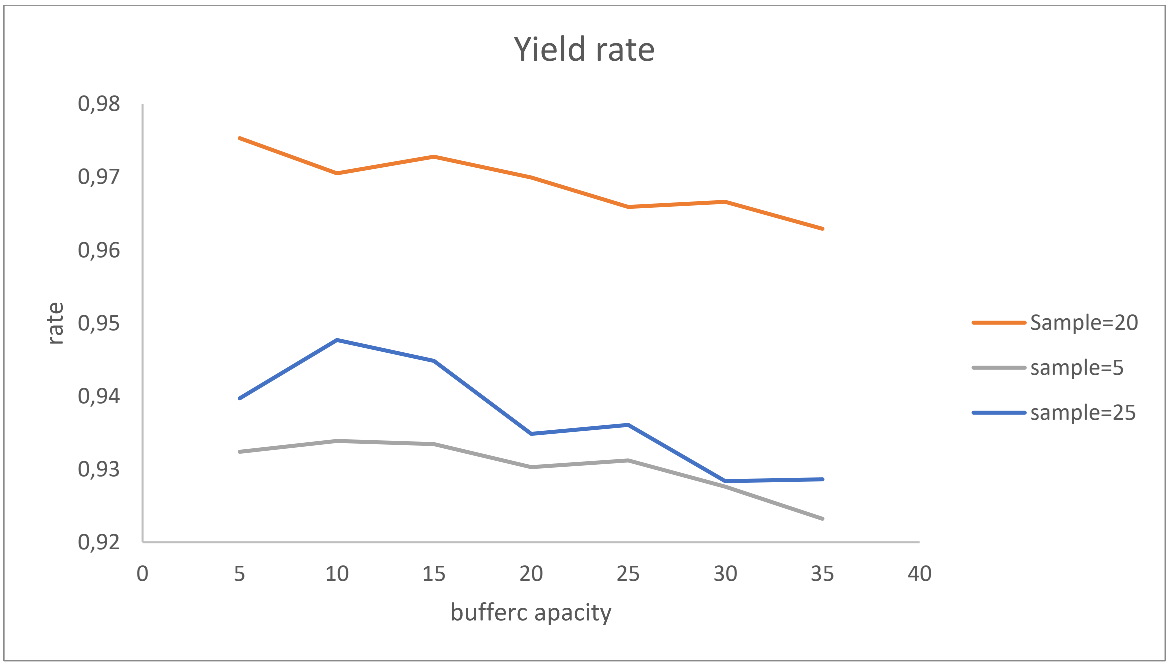


Figure 8.13: sensitivity differences of yield with different sample size configuration

8.1.3. Machine service rate variation

If upstream machine speed is increased keeping the same speed of inspection, the performance of yield rate production will increase ass well (Fig.8.16) . In fact, whenever the machine speeds up, the buffer is would be always full. Consequently, according to how the degradation is designed in simulation, if the machine is kept under usage, with lower time execution for a single wafer it will deteriorated slowly. Therefore, it would produce deviated product less frequently, and generate less number of intervention (Fig.8.17). This phenomenon is confirmed by the comparison of same sensitivity analysis among two different Buffer capacity (Fig.8.18). For a small buffer, a higher speed tends to fulfil easily the buffer and to keep frequently the upstream into idle state.

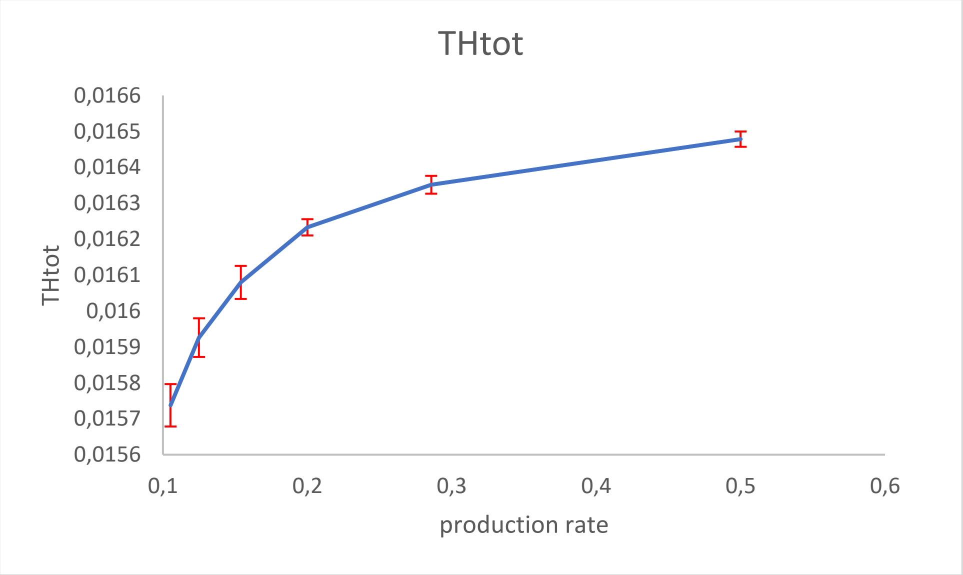


Figure 8.14 Throughput total by varying upstream machine speed

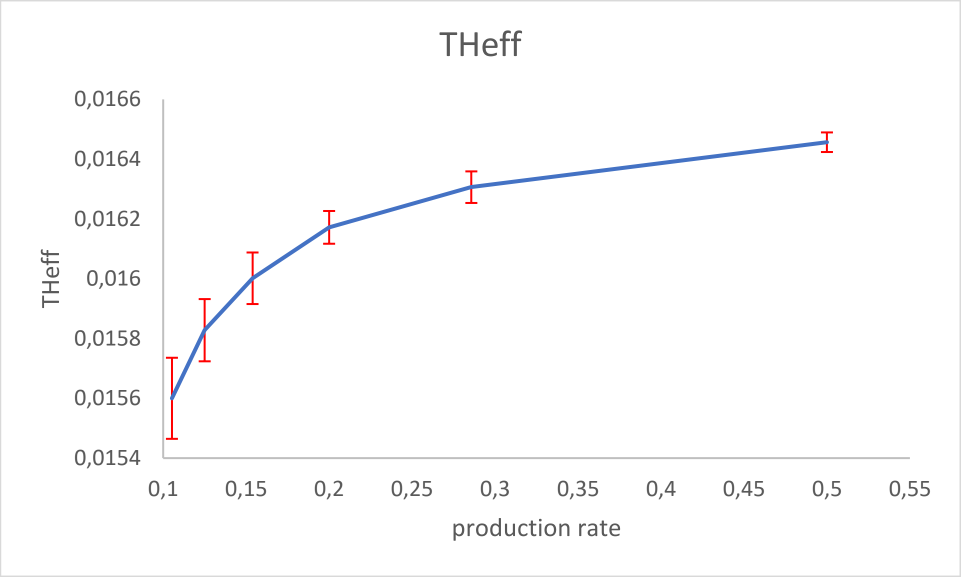


Figure 8.15 Throughput total by varying upstream speed

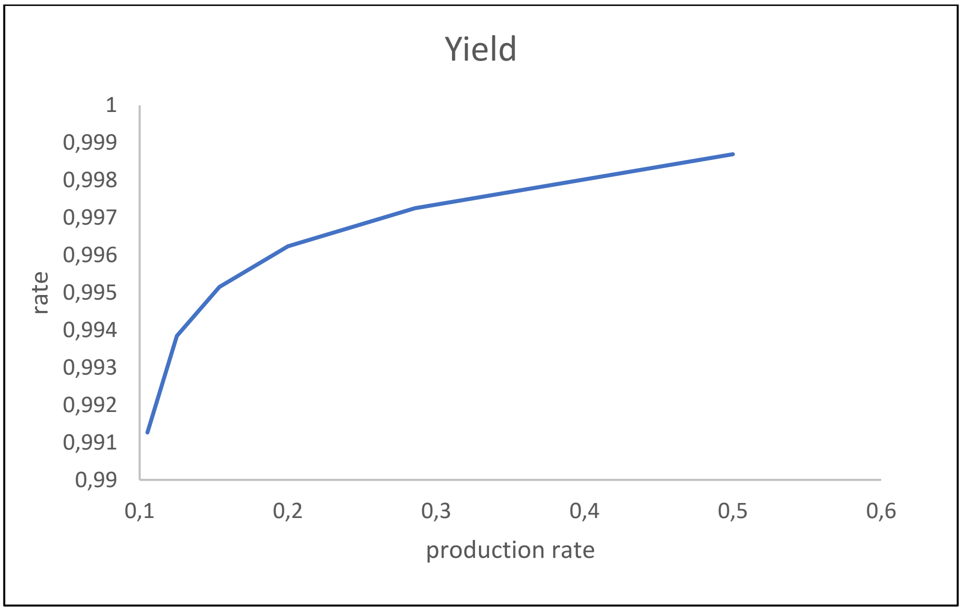


Figure 8.16 Yield rate by varying upstream machine speed

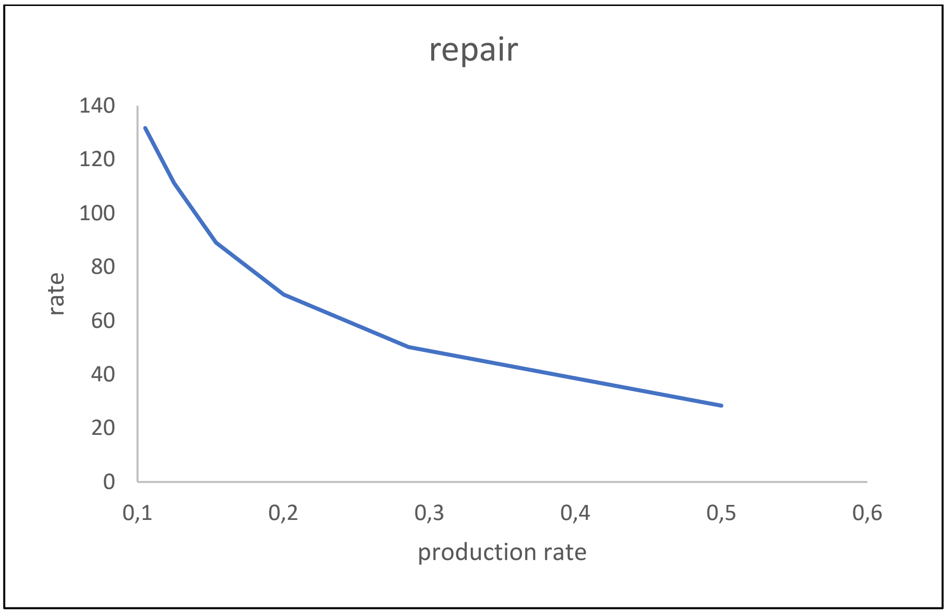


Figure 8.17 number of repairs by upstream machine speed up

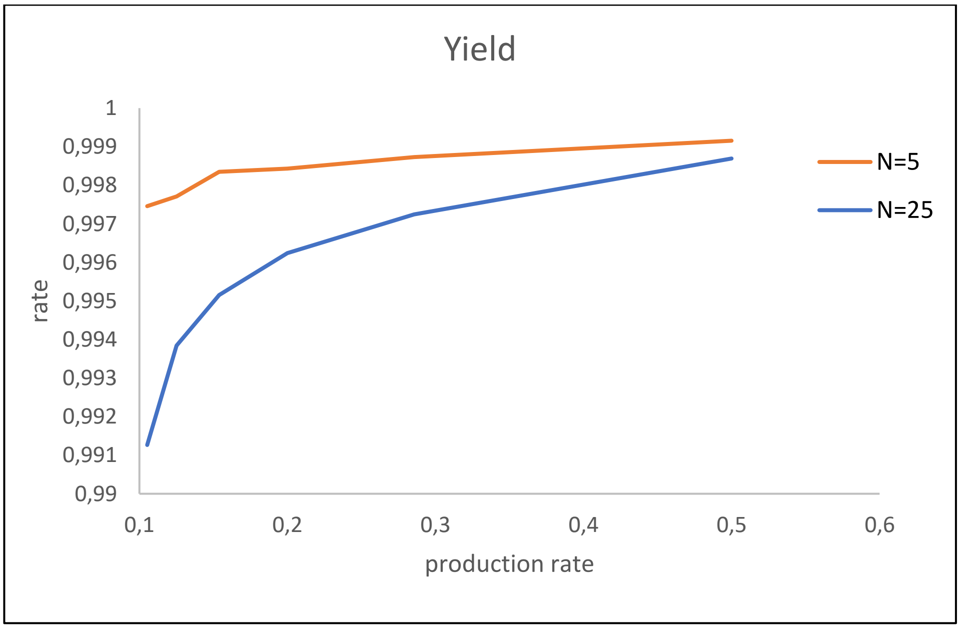
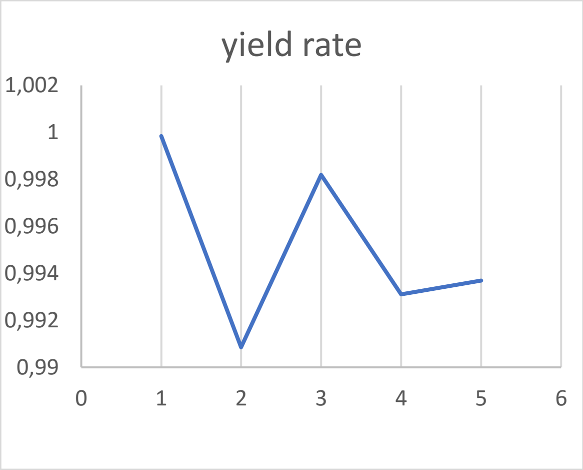
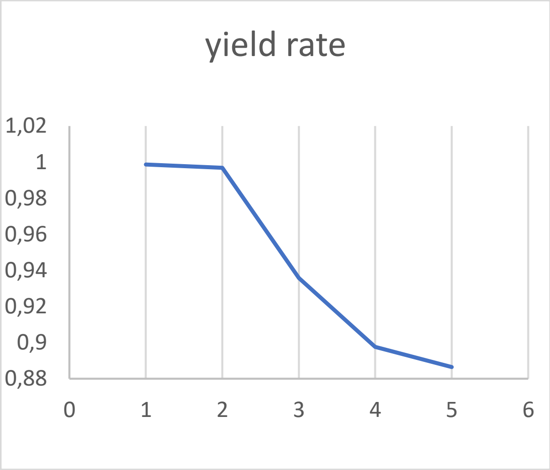
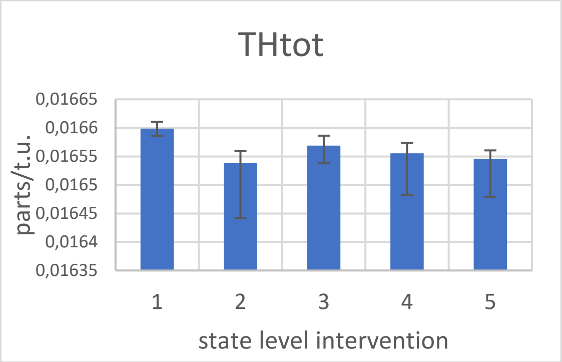
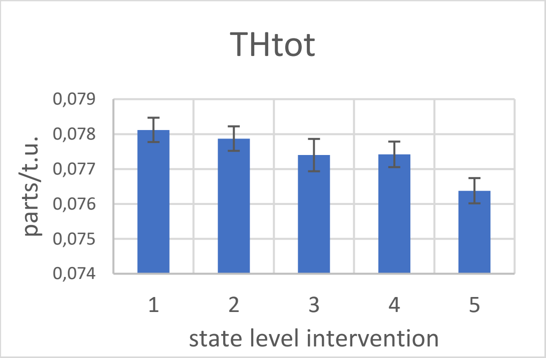
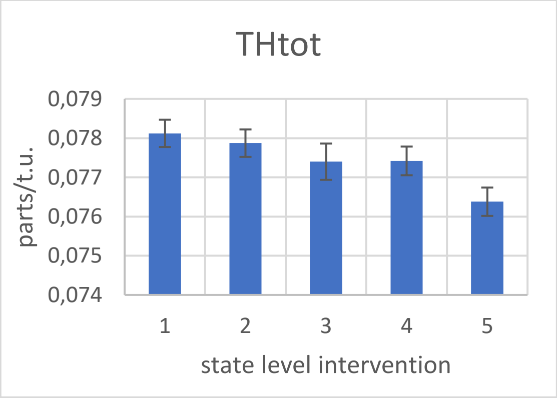
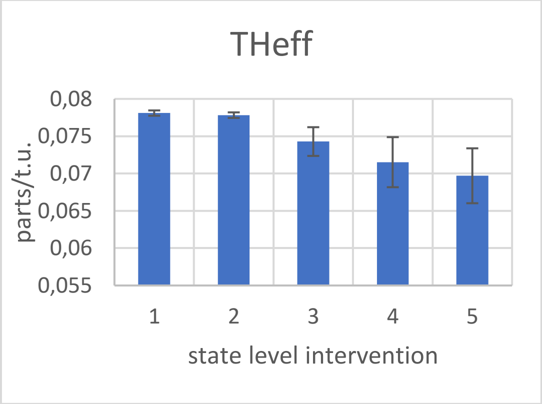


Figure 8.18 Yield rate sensitivity comparison between different buffer capacity

8.1.4. Machine control sensitivity

The next analysis strongly motivates the objective of this thesis for the proper application of APC and SPC joint control. As a matter of fact, the decision of Maintenance policy applied on the state equipment condition estimates is strongly affected by the system configurations. For example, one important parameter that affect the policy decision making done from in situ machine station is the sampling fraction inspected in the downstream machine. Figures, show that the decision-making of the state to intervene for prevention changes along with the sampling fraction. High productivity can be attained at the expense of yield production, if we use the lowest sampling configuration as an example. As a matter of fact , the low sampling boosts production in isolation from the inspection station, delivering the product waiting in the buffer more quickly. This entails less blocking conditions and more frequent use of the upstream machine, which deteriorates more quickly with time. Furthermore, waiting until the very end of the predicted degradation to intervene is not advised because low inspection frequency increases the probability of properly detecting upstream degradation. Therefore, it is more effective and convenient to make decisions within the early stages of interventions.

In contrast, if we examine a sampling fraction of 100%, we can see that states 1 and 2 might no longer be the best option. This is likely because a higher sampling makes a more detailed analysis of the deviations, which makes it more practical to stop the machine later, and a lower sampling will result in the upstream machine being offline more frequently.

Sample size= 25	Sample size= 5
 	
 	
 	

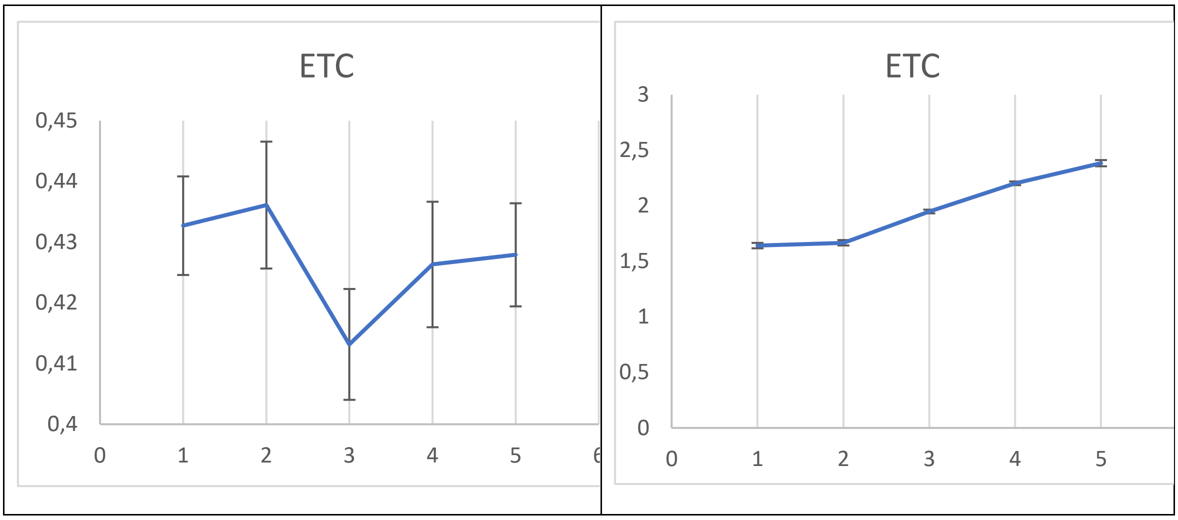


Figure 8.19 Comparative study of performances with different sample size

8.1.5. Quality control

The system's design also affects how decisions under the CBM quality control policy are made. A threshold limit that is closer to the imposed upper limit control of EWMA chart results in yield losses because the policy kicks in while the deterioration is already well along the way, inevitably harming the product's final quality. This effect weighs more when a greater buffer design is applied, as can be shown in the Fig.8.20 below. As a result, while quality maintenance performed as late as possible reduces the number of preventive interventions, an higher number of corrective interventions might result in lengthy system downtimes.

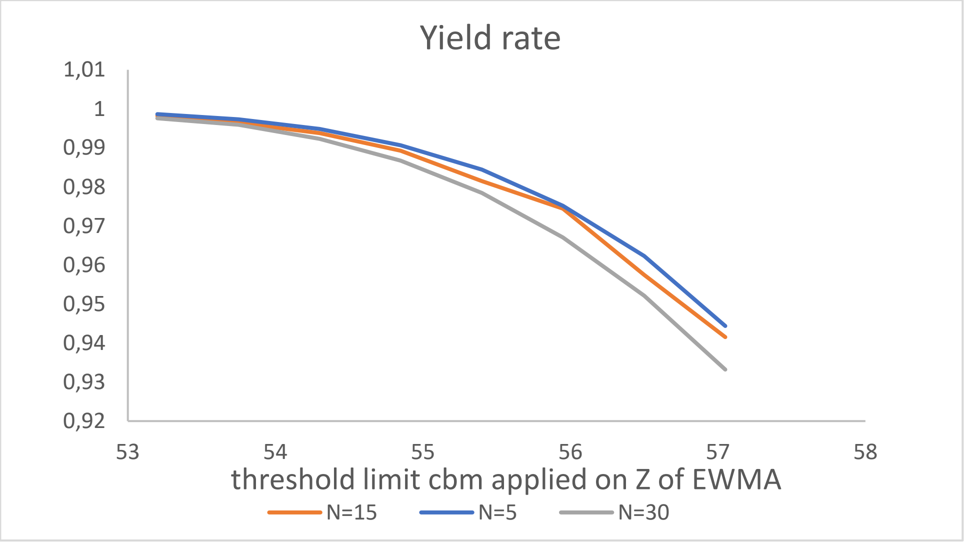


Figure 8.20 Yield rate comparisons among different buffer capcacity

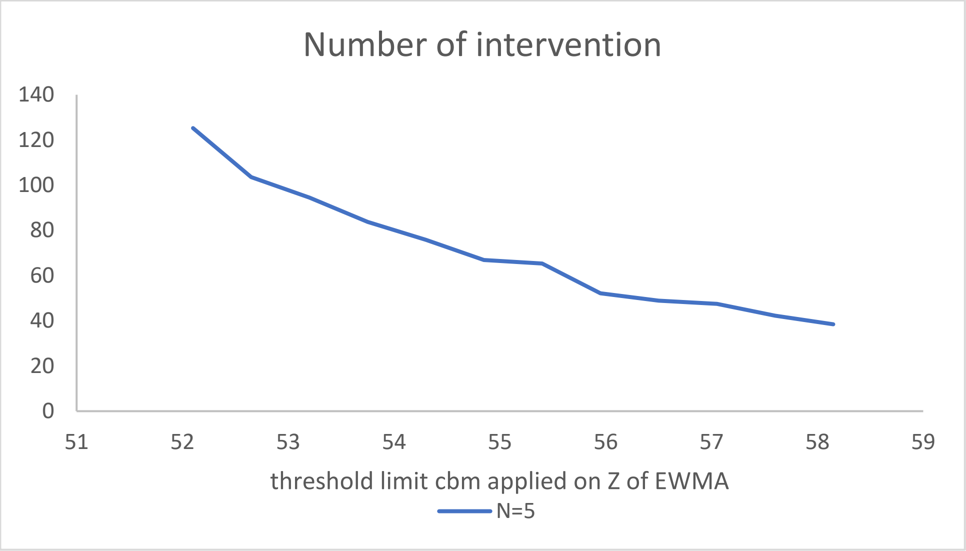


Figure 8.21 number of preventive interventions dictated by the quality control

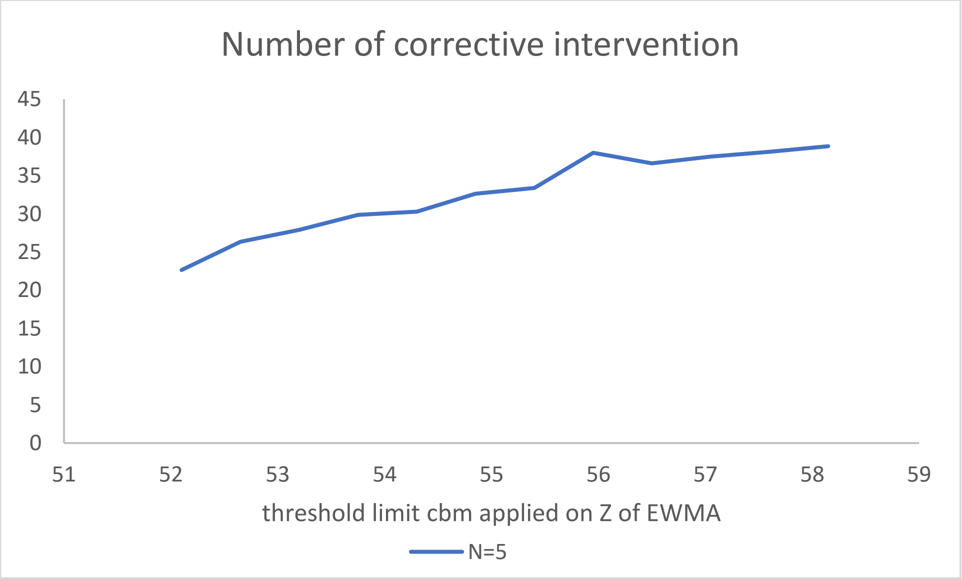


Figure 8.22 number of corrective interventions

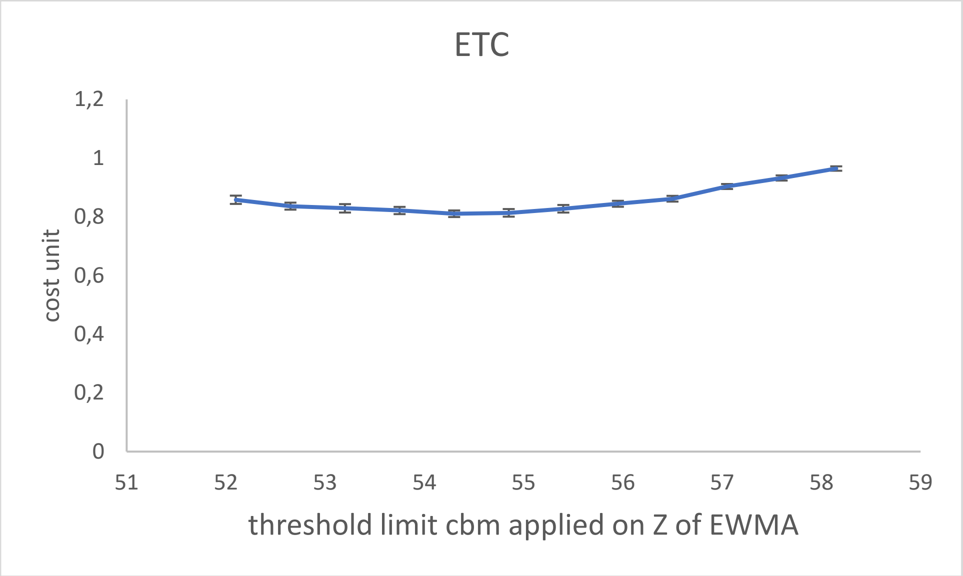


Figure 8.23 Expected total cost

8.2. Response surface analysis

8.2.1. Response surface subject to buffer variation

This section explores the implementation of various policy threshold combinations using DOE full factorial and inscribed campaign designs. The polynomial regression is achieved for the relevant Yield, Throughput effective, and projected Total Cost.

In order to comprehend how the trade-off between the two policy threshold varies with the design modification of a single manufacturing system configuration, a Comparative analysis of multiple response surfaces is done before delving further into the optimization.

First, the configuration design's buffer capacity is examined. As can be seen, the isocurves behave similarly for the three configurations of the buffer capacity (N=25, N=50, and N=75).

In fact, the following synergies of machine maintenance and quality control result in the same yield and THeff performances: Preventive intervention at the beginning of the predicted deterioration process by the machine controller with a cbm threshold set near to the top control limit might be one possibility. In this method, machine control takes the lead in assessing preventative intervention, whereas CBM quality control plays a minor role in delivering feedback concerning significant quality deviations that the machine controller has not been able to stop.

Otherwise, a different solution strategy is to keep the weighted exponential function's limit, which is traced run by run, far from the LCU so that it can be activated in case of small deviations or once the machine controller notices that the machine has reached the last estimated state of machine degradation.

It is unclear whether to give the machine controller priority, when to use CBM control for remote detection, or even when to create a balance between the two in order to achieve the best "good production" due to the manufacturing system configuration's constant variation. The tradeoffs identified by the isocurve shift toward early phases and a lower CBM threshold for an expanding buffer capacity. It can be rationalized by the fact that the inter operational buffer's strong decoupling effect causes a larger feedback latency, making it unable to respond to controls at higher threshold levels. According to how the costs of preventive maintenance, corrective maintenance, scrap cost and production cost had been structured, the cheapest solution fosters the predominance of the machine control feedback against the quality control one. Naturally this last statement can change a lot if the structure of cost gives different importance to preventive interventions or scrap efforts. Moreover, the final optimization should take into account also cost related to outsourcing production and investment of sensoring device to apply over the control machine, which can change definitely the strategical policy of the firm.

8.2.2. Response surface subject to service rate variation of upstream machine

The design of various upstream operation execution speeds has a greater influence on policy decision-making. The tradeoff is the same as that outlined for the investigation of the variation in buffer capacity, as shown in the Figures. By speeding up the upstream machine in this instance as well, it is feasible to see how the policy choice changes. In fact, the most effective approach for slower machine operations appears to be to give attention to the first stages of degradation detected by the machine controller and to notify the system after significant deviations have not been avoided by estimates from the machine controller. This probably happens because the machine is free to serve the buffer with lower service rate with lower probability of blocking. As a result the machine usage increase as well as the rhythm of degradation. Since there is only one lot of buffer capacity in this study, a significant level of delay hinder the quality control. A rapid decline must therefore be detected as soon as possible. The upstream machine, however, tends to be kept underutilized and frequently in a blocking state for greater service rates. This results in a more gradual decline. As a matter of fact, the quality control can be connected with the machine controller to read slower deviations of the parameter that was measured and to achieve good yield and TH eff performances. Additionally, when upstream equipment gains speed, the cost curve's shape changes.

8.2.3. Response surface subject to changes of sample size

Finally, the response surfaces of the comparative analysis to modification in sample fraction are investigated. As the sample fraction rises, the decision-making changes. The management of deviation offers great responsiveness in identifying deviations with high sampling.

Additionally, a higher sample rate may result in a reduced usage of the upstream machine, which frequently enters an idle state. This enables the deterioration process to be slowed down. Therefore, compared to the scenario of having a low sampling control, the quality control might have a higher priority with a high sample rate.

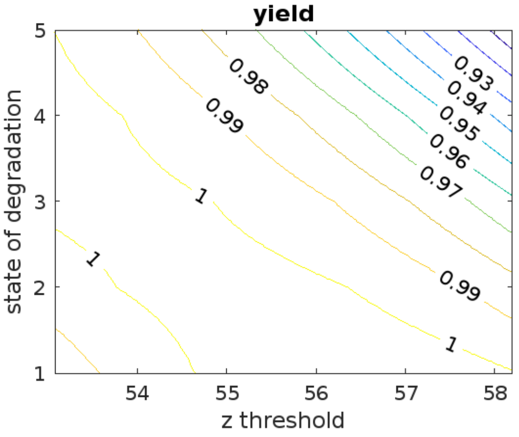
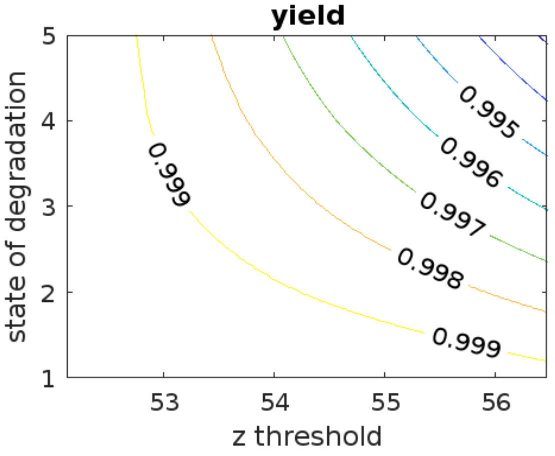
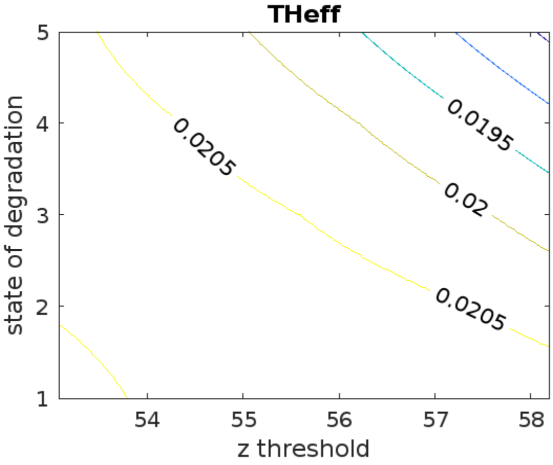
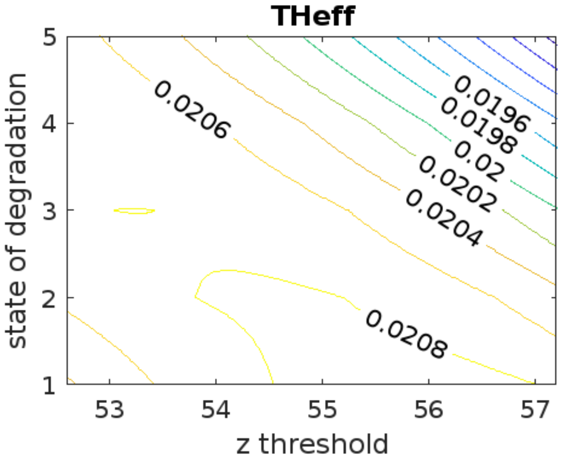
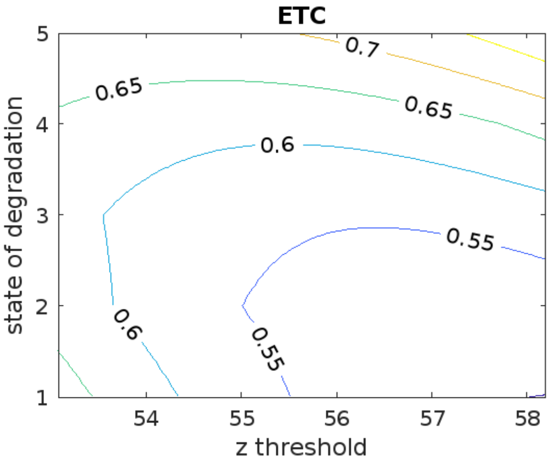
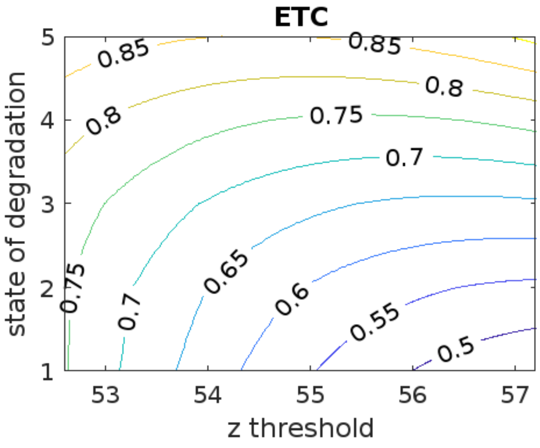
N=25 wafers, 1 lot	N=75 wafers , 3 lots
	
	
	

Figure 8.24 Response surfaces comparisons between two different buffer size configuration

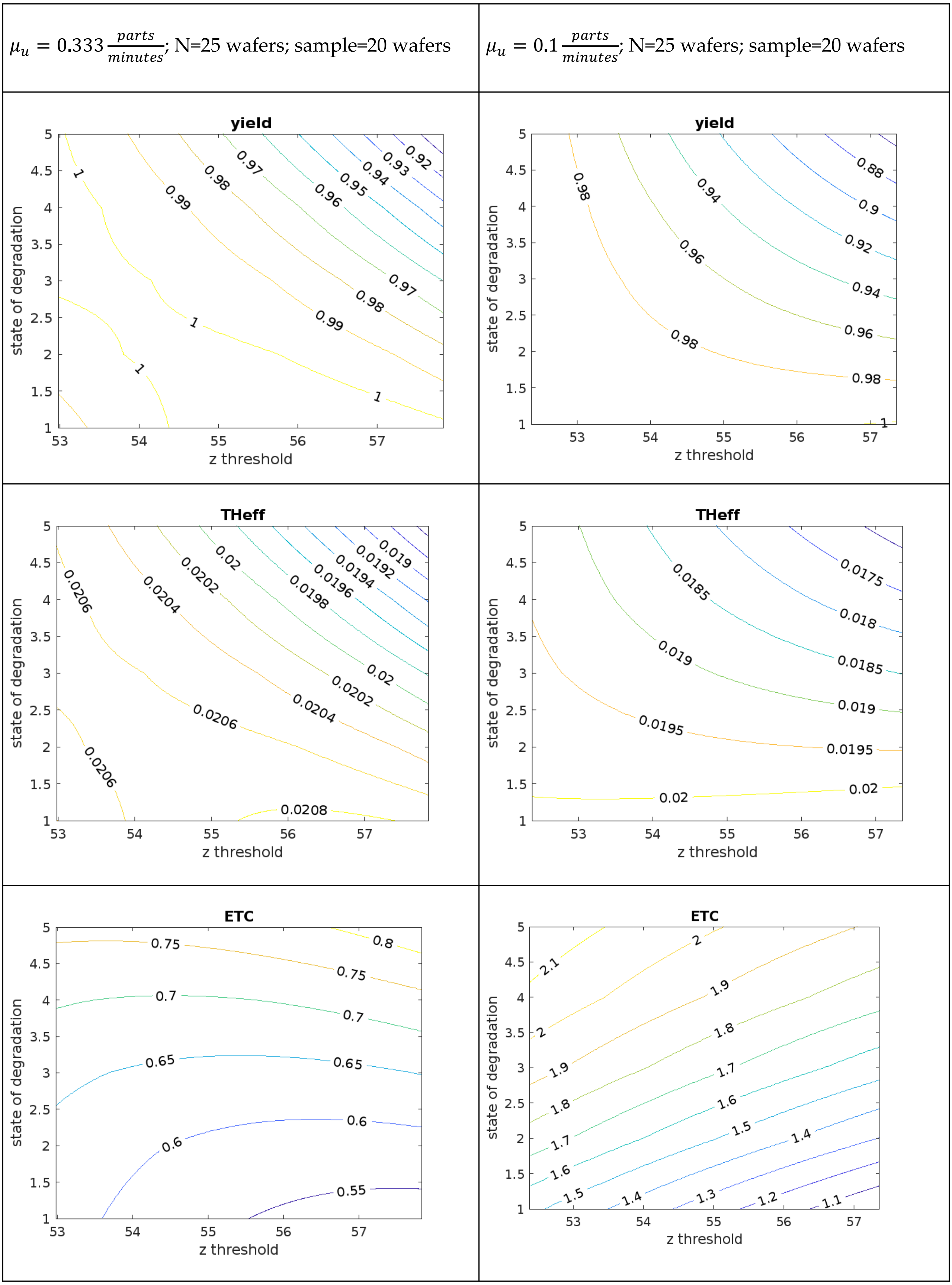


Figure 8.25 Response surfaces comparisons between two different upstream machine speed

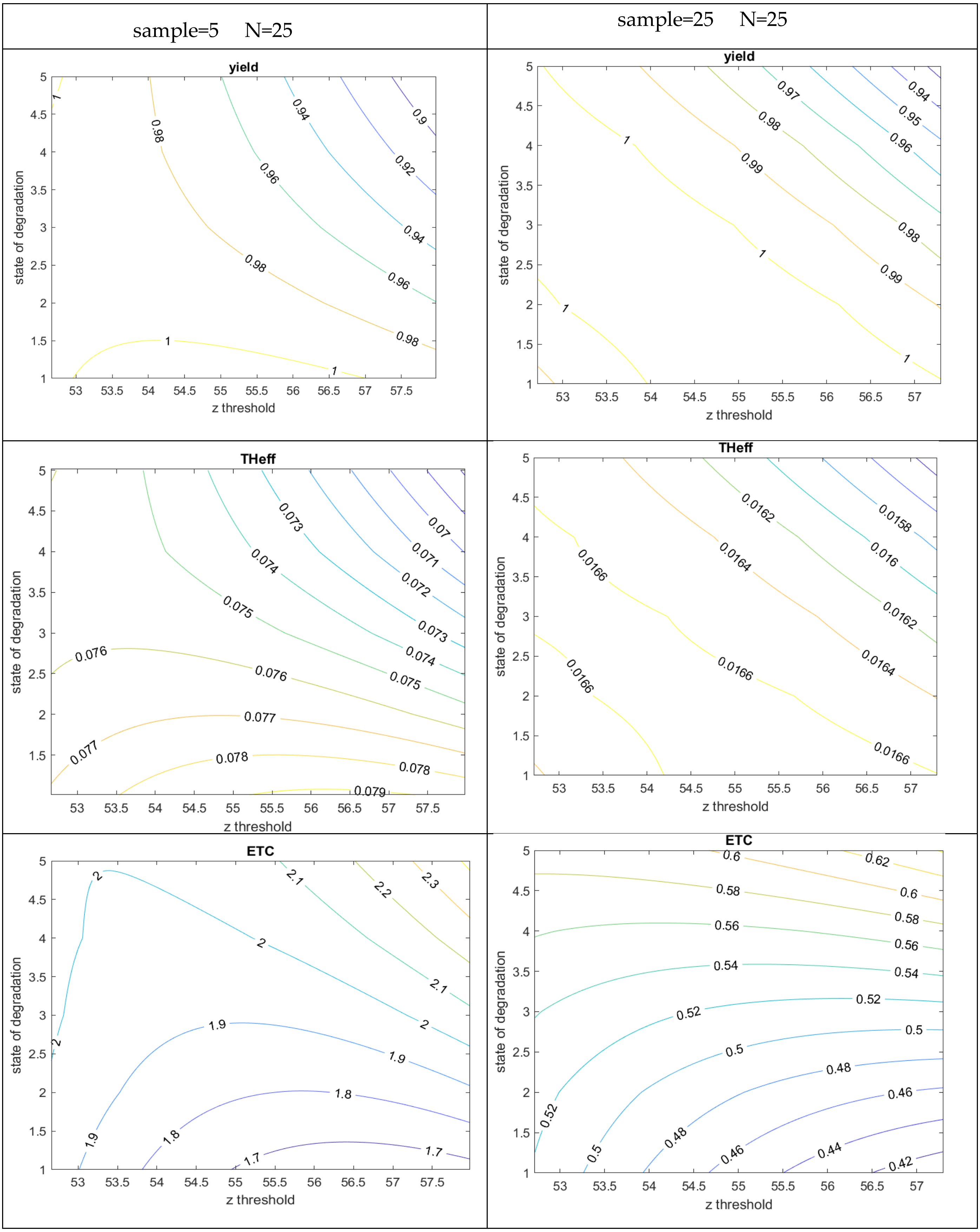


Figure 8.26 Response surfaces comparisons between two different sample size configuration

8.3. Final optimization result

The first problem of optimization is formulated in model 8.1:

$$\begin{cases} Max \quad Yield(Machine\ threshold, EWMA\ threshold) & (8.1a) \\ \quad\quad Subject\ to & \\ l_{low} \leq Machine\ threshold \leq l_{high} & (8.2b) \\ L_{low} \leq EWMA\ threshold \leq L_{high} & (8.3c) \\ Machine\ threshold\ \in N & (8.4d) \\ Ewma\ threashold \in R & (8.5e) \end{cases}$$

(8. 6a) Maximization of the Yield

(8. 7b)$Machine\ threshold\ must\ be\ within\ the\ limit\ admitted\ designed\ in\ DOE$ (l_{low}, l_{high}).

(8. 8c)$Quality\ threshold\ must\ be\ within\ the\ limit\ designed\ in\ DOE\ (L_{high}\ , L_{low})$.

(8. 9d) Machine threshold is a discrete variable

(8. 10e) Machine threshold is a continuous variable

The second problem formalization aimed to maximize the Yield rate is similar to 8.1. It is represented in 8.2 below:

$$\begin{cases} Max \quad THeff(Machine\ threshold, EWMA\ threshold) \\ Subject\ to \\ \quad l_{low} \leq Machine\ threshold \leq l_{high} \\ \quad L_{low} \leq EWMA\ threshold \leq L_{high} \\ \quad Machine\ threshold\ \in N \\ \quad Ewma\ threashold \in R \end{cases} \quad (8.11)$$

The final optimization task involves minimizing the overall total cost incurred to meet a specified demand level (d) within a specified time frame (T). the formulation of the problem is described below (8.3).

$$\begin{cases} Min \quad Total\ Cost = ETC(.) \times T + (d - THeff(.) \times T) \times C_{inventory} & (8.3a) \\ \quad Subject\ to & \\ \quad\quad THeff * T \geq d & (8.3b) \\ \quad\quad ETC \geq 0 & (8.3c) \\ \quad\quad l_{low} \leq Machine\ threshold \leq l_{high} & (8.3d) \\ \quad\quad L_{low} \leq EWMA\ threshold \leq L_{high} & (8.3e) \\ \quad\quad Machine\ threshold\ \in N & (8.3f) \\ \quad\quad Ewma\ threashold \in R & (8.3g) \end{cases}$$

(8. 3a) Minimization of the overall cost

(8. 3b)the production capacity must satisfy the demand in the time horizon T

(8. 3c)$the\ cost\ function\ of\ the\ respons\ surface\ must\ be\ positive$

(8. 3d)$Machine\ threshold\ must\ be\ within\ the\ limit\ admitted\ designed\ in\ DOE$ (l_{low}, l_{high}).

(8. 3e)$Quality\ threshold\ must\ be\ within\ the\ limit\ designed\ in\ DOE\ (L_{high}\ , L_{low})$.

(8. 3f) Machine threshold is a discrete variable

(8. 3g) Machine threshold is a continuous variable

In this section the optimization of every response surface (RS) extracted for every configuration is addressed. Three optimization perspective are performed: maximization of effective throughput Table 8.1, minimization, the maximization of the yield rate Table 8.2 and the minimization of the expected total cost Table 8.3.

Table 8.1: optimization result of RS by varying buffer size

max (Yield)			
Buffer size	Threshold cbm	Threshold PM	yield
5	54,5	1	1
25	51	5	1
50	55,6	1	1
75	51	5	1
max (Theff)			
Buffer size	Threshold cbm	Threshold PM	Throughput effective
5	56,26	1	0,020563961
25	54,79	1	0,020689713
50	51,10	4	0,020772578
75	54,75	1	0,02083459
min (ETC)			
Buffer size	Threshold cbm	Threshold PM	ETC
25	51	5	11420,76372
50	55,87	1	6732,272652
75	51	5	14757,33964
5	54,7	1	6783,129975

Table 8.2: optimization result of RS by varying service time

max (Theff)			
Speed	Threshold cbm	Threshold PM	Throughput effective
1	55	1	0,020766
1,5	58	1	0,020794
3	54,5	1	0,020707
6	56,5	1	0,020549
10	57,3	1	0,020461

max (Yield)			
Speed	Threshold cbm	Threshold PM	yield
1	54,27	1	1
1,5	51	5	1
3	51	5	1
6	51	5	1
10	51	5	1

min (ETC)			
speed	Threshold cbm	Threshold PM	ETC
1	55,09	1	5682,231
1,5	51	5	6530,036
3	51	5	12092,4
6	55,77	1	12387,71
10	58	1	14975,02

Table 8.3: optimization result of RS by varying sample size

max (Theff)			
Sample size	Threshold cbm	Threshold PM	Throughput effective
5	55,087	1	0,020766
10	58	1	0,020794
15	54,51	1	0,020707
20	56,55	1	0,020549
25	57,33	1	0,020461

max (Yield)			
Sample size	Threshold cbm	Threshold PM	yield
5	54,27	1	1
10	51	5	1
15	51	5	1
20	51	5	1
25	51	5	1

min (ETC)			
Sample size	Threshold cbm	Threshold PM	ETC
5	55,09	1	5682,231
10	51	5	6530,036

15	51	5	12092,4
20	55,75	1	12408,73
25	58	1	14975,02

9 Conclusion

In this thesis, a discrete event simulation model is developed to explain an advanced control loop that is proposed for an asynchronous production line with two machines and one buffer. The simulation attempts to replicate the dynamics of a semiconductor production line where several wafer fabrication processes may lead to unrecognizable degradation states that affect the output quality. In this simulation, the upstream machine's degradation phenomena is tracked locally by machine control and remotely by SPC inspection control through an EWMA chart. The integrated control policies seek to find a solution to lower the output of scrap and improve the quality of production. The model simulates a prognostic machine intervention which is made possible by the placement of an in-situ machine controller. This controller employs an HMM model to keep track of a roughly stratified function of the actual state of equipment deterioration that is not readily observable. On inspection side, the quality SPC control is implemented. It provides feedback about product quality deviation through a run-to-run control of the EWMA control chart. The simulation replicates this loop and serves as a tool to comprehend the key performance indicators for the zero-defect manufacturing paradigm, such as throughput efficiency and yield rate. The objective of this thesis is to understand how machine controller policy and quality control policy interact and how the policy decision making is affected by changing structure of the manufacturing system. First, a sensitivity analysis has been conducted using a simulation model to demonstrate how the inter-operational buffer size, sample size, and machine speeds significantly impact yield and throughput performance as well as the appropriate policy thresholds. A response surface methodology is used to illustrate the trade-off between the two threshold controls and how this trade-off varies with large changes in the manufacturing system design. This strategy aims to roughly reflect how the system's performance analysis, in response to various policy changes, behaves given a range of manufacturing system setting. Afterwards, short analysis of variance and an evaluation of the residuals dispersion of each regression function is performed to confirm every response surface fitting. Finally, a comparison study of various outcomes with different configurations is conducted. The end findings show that when machine maintenance and quality control are considered in a single framework, system engineering prioritizes the visibility of the two control systems differently, resulting to a different optimal solution.

9.1. Further Research

The thesis fits in the research area of simulation and metamodeling optimization of manufacturing systems. It takes into account the necessity of coming up with a joint policy of quality and machine maintenance control within a special framework. Therefore, there are a lot of opportunities for new advancements in the future. Some have a tight connection to the suggested model:

- Simulation model validation through the creation of an analytical tool or by the use of a real-world case study
- Extension to multi stage manufacturing system remotely monitored
- Introduction of more wafer recipes introduced in the system

9.2. Bibliography and citations

[1] «Growing at a slower pace, world population is expected to reach 9.7 billion in 2050 and could peak at nearly 11 billion around 2100 | UN DESA | United Nations Department of Economic and Social Affairs». https://www.un.org/development/desa/en/news/population/world-population-prospects-2019.html (consultato 3 aprile 2022).

[2] «Evidence of decoupling consumption-based CO2 emissions from economic growth | Elsevier Enhanced Reader». https://reader.elsevier.com/reader/sd/pii/S2666792421000664?token=37378DF0158BBC6F233AF87B60DCDF4B0886FB11E70D39621D172E462E1C8FB9F53A45851DFF129DFEF15B250C891B53&originRegion=eu-west-1&originCreation=20220608125420 (consultato 8 giugno 2022).

[3] Martin, «Sustainable consumption and production», *United Nations Sustainable Development*. https://www.un.org/sustainabledevelopment/sustainable-consumption-production/ (consultato 8 giugno 2022).

[4] A. Matta, «Manufacturing Strategy», pag. 60.

[5] «THE 17 GOALS | Sustainable Development». https://sdgs.un.org/goals (consultato 4 aprile 2022).

[6] A. B. Horin, «Equipment maintenance policies impact on Flow Time and Yield given inspection policy», pag. 68, 2012.

[7] T. Tolio, D. M. C. Magnanini, e D. Djurdjanovic, «Performance evaluation of inspection policies in semiconductor fabrication», pag. 233.

[8] «Introduction to semiconductor technology», pag. 15.

[9] M. Colledani, «OPTIMIZATION OF INSPECTION STATION ALLOCATION IN SERIAL MANUFACTURING LINES», pag. 100.

[10] M. C. Magnanini e T. Tolio, «Switching- and hedging- point policy for preventive maintenance with degrading machines: application to a two-machine line», *Flex. Serv. Manuf. J.*, vol. 32, n. 2, pagg. 241–271, giu. 2020, doi: 10.1007/s10696-019-09370-7.

[11] T. A. M. Tolio e A. Ratti, «Performance evaluation of two-machine lines with generalized thresholds», *Int. J. Prod. Res.*, vol. 56, n. 1–2, pagg. 926–949, gen. 2018, doi: 10.1080/00207543.2017.1420922.

[12] M. Colledani *et al.*, «Design and management of manufacturing systems for production quality», *CIRP Ann.*, vol. 63, n. 2, pagg. 773–796, 2014, doi: 10.1016/j.cirp.2014.05.002.

[13] J. L. Callen, C. Fader, e I. Krinsky, «Just-in-time: A cross-sectional plant analysis», *Int. J. Prod. Econ.*, vol. 63, n. 3, pagg. 277–301, gen. 2000, doi: 10.1016/S0925-5273(99)00025-0.

[14] J. Owen e D. Blumenfeld, «Effects of operating speed on production quality and throughput», *Int. J. Prod. Res.*, vol. 46, pagg. 7039–7056, dic. 2008, doi: 10.1080/00207540701227833.

[15] N. Davari, B. Veloso, G. de A. Costa, P. M. Pereira, R. P. Ribeiro, e J. Gama, «A Survey on Data-Driven Predictive Maintenance for the Railway Industry», *Sensors*, vol. 21, n. 17, pag. 5739, ago. 2021, doi: 10.3390/s21175739.

[16] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, e G. Elger, «Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry», *Reliab. Eng. Syst. Saf.*, vol. 215, pag. 107864, nov. 2021, doi: 10.1016/j.ress.2021.107864.

[17] A. Kanawaday e A. Sane, «Machine learning for predictive maintenance of industrial machines using IoT sensor data», in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, nov. 2017, pagg. 87–90. doi: 10.1109/ICSESS.2017.8342870.

[18] F. Arena, M. Collotta, L. Luca, M. Ruggieri, e F. G. Termine, «Predictive Maintenance in the Automotive Sector: A Literature Review», *Math. Comput. Appl.*, vol. 27, n. 1, Art. n. 1, feb. 2022, doi: 10.3390/mca27010002.

[19] Y. Liu, «PREDICTIVE MODELING FOR INTELLIGENT MAINTENANCE IN COMPLEX SEMICONDUCTOR MANUFACTURING PROCESSES», pag. 167.

[20] G. A. Susto, S. Pampuri, A. Schirru, G. D. Nicolao, S. McLoone, e A. Beghi, «Automatic Control and Machine Learning for Semiconductor Manufacturing: Review and Challenges», pag. 8, 2012.

[21] S. Munirathinam, D. B. Ramadoss, e M. Tech, «Big Data Predictive Analtyics for Proactive Semiconductor Equipment Maintenance», pag. 10.

[22] S. Sahoo e S. Yadav, «Influences of TPM and TQM Practices on Performance of Engineering Product and Component Manufacturers», *Procedia Manuf.*, vol. 43, pagg. 728–735, 2020, doi: 10.1016/j.promfg.2020.02.111.

[23] A. Al-Refaie, N. Lepkova, e M. E. Camlibel, «The Relationships between the Pillars of TPM and TQM and Manufacturing Performance Using Structural Equation Modeling», *Sustainability*, vol. 14, n. 3, pag. 1497, gen. 2022, doi: 10.3390/su14031497.

[24] L. M. Wein, «On the relationship between yield and cycle time in semiconductor wafer fabrication», *IEEE Trans. Semicond. Manuf.*, vol. 5, n. 2, pagg. 156–158, mag. 1992, doi: 10.1109/66.136277.

[25] I. Tirkel, N. Reshef, e G. Rabinowitz, «In-line Inspection Impact on Cycle Time and Yield», *IEEE Trans. Semicond. Manuf.*, vol. 22, n. 4, pagg. 491–498, nov. 2009, doi: 10.1109/TSM.2009.2031779.

[26] M. Bureau, S. Dauzère-Pérès, e Y. Mati, «SCHEDULING CHALLENGES AND APPROACHES IN SEMICONDUCTOR MANUFACTURING», *IFAC Proc. Vol.*, vol. 39, n. 3, pagg. 739–744, 2006, doi: 10.3182/20060517-3-FR-2903.00370.

[27] «Delayed Response – A Red Flag in Manufacturing», *Parsable*, 30 ottobre 2020. https://parsable.com/blog/quality/delayed-response-a-red-flag-in-manufacturing/ (consultato 13 maggio 2022).

[28] M. C. Magnanini, M. Colledani, e D. Caputo, «Reference architecture for the industrial implementation of Zero-Defect Manufacturing strategies», *Procedia CIRP*, vol. 93, pagg. 646–651, 2020, doi: 10.1016/j.procir.2020.05.154.

[29] M. Rezaei-Malek, M. Mohammadi, J.-Y. Dantan, A. Siadat, e R. Tavakkoli-Moghaddam, «A review on optimisation of part quality inspection planning in a multi-stage manufacturing system», *Int. J. Prod. Res.*, vol. 57, n. 15–16, pagg. 4880–4897, ago. 2019, doi: 10.1080/00207543.2018.1464231.

[30] H. Rivera-Gómez, A. Gharbi, J.-P. Kenné, O. Montaño-Arango, e J. R. Corona-Armenta, «Joint optimization of production and maintenance strategies considering a dynamic sampling strategy for a deteriorating system», *Comput. Ind. Eng.*, vol. 140, pag. 106273, feb. 2020, doi: 10.1016/j.cie.2020.106273.

[31] M. Mohammadi, «Mathematical modelling of a robust inspection process plan: Taguchi and Monte Carlo methods», *Int. J. Prod. Res.*, pag. 24.

[32] M. REZAEI-MALEK, R. Tavakkoli-Moghaddam, A. Siadat, e J.-Y. Dantan, «A novel model for the integrated planning of part quality inspection and preventive maintenance in a linear-deteriorating serial multi-stage manufacturing system», *Int. J. Adv. Manuf. Technol.*, vol. 96, n. 9–12, pagg. 3633–3650, 2018, doi: 10.1007/s00170-018-1751-1.

[33] M. Mohammadi, J.-Y. Dantan, A. Siadat, e R. Tavakkoli-Moghaddam, «A bi-objective robust inspection planning model in a multi-stage serial production system», *Int. J. Prod. Res.*, vol. 56, n. 4, pagg. 1432–1457, ago. 2017, doi: 10.1080/00207543.2017.1363425.

[34] M. Rezaei-Malek, A. Siadat, J.-Y. Dantan, e R. Tavakkoli-Moghaddam, «An Approximation Approach for an Integrated Part Quality Inspection and Preventive Maintenance Planning in a Nonlinear Deteriorating Serial Multi-stage

Manufacturing System», *IFAC-Pap.*, vol. 51, n. 11, pagg. 270–275, 2018, doi: 10.1016/j.ifacol.2018.08.291.

[35] M. Colledani e T. Tolio, «Integrated analysis of quality and production logistics performance in manufacturing lines», *Int. J. Prod. Res.*, vol. 49, n. 2, pagg. 485–518, gen. 2011, doi: 10.1080/00207540903443246.

[36] B. Bouslah, A. Gharbi, e R. Pellerin, «Joint production, quality and maintenance control of a two-machine line subject to operation-dependent and quality-dependent failures», *Int. J. Prod. Econ.*, vol. 195, pagg. 210–226, gen. 2018, doi: 10.1016/j.ijpe.2017.10.016.

[37] M. C. Magnanini e T. Tolio, «Restart policies to maximize production quality in mixed continuous-discrete multi-stage systems», *CIRP Ann.*, vol. 69, n. 1, pagg. 361–364, 2020, doi: 10.1016/j.cirp.2020.03.021.

[38] K. Okamura e H. Yamashina, «Analysis of the Effect of Buffer Storage Capacity in Transfer Line Systems», *E Trans.*, vol. 9, n. 2, pagg. 127–135, giu. 1977, doi: 10.1080/05695557708975134.

[39] M. A. Jafari e J. G. Shanthikumar, «An Approximate Model of Multistage Automatic Transfer Lines with Possible Scrapping Of Workpieces», *IIE Trans.*, vol. 19, n. 3, pagg. 252–265, set. 1987, doi: 10.1080/07408178708975394.

[40] «A review of: "Stochastics Models of Manufacturing Systems" JOHN A. BUZACOTT, J. GEORGE SHANTHIKUMAR, 1993 Englewood Cliffs, NJ, Prentice Hall ISBN 0 13 847567 9 $84.20», *Eur. J. Eng. Educ.*, vol. 18, n. 2, pagg. 218–219, gen. 1993, doi: 10.1080/03043799308928355.

[41] G. Liberopoulos, G. Kozanidis, e P. Tsarouhas, «Performance Evaluation of an Automatic Transfer Line with WIP Scrapping During Long Failures», *Manuf. Serv. Oper. Manag.*, vol. 9, n. 1, pagg. 62–83, gen. 2007, doi: 10.1287/msom.1060.0118.

[42] J. Kim e S. B. Gershwin, «Integrated Quality and Quantity Modeling of a Production Line», pag. 32.

[43] F. Trojan e R. F. M. Marçal, «Proposal of Maintenance-types Classification to Clarify Maintenance Concepts in Production and Operations Management», pag. 14.

[44] K. Fraser, «Facilities management: the strategic selection of a maintenance system», *J. Facil. Manag.*, vol. 12, n. 1, pagg. 18–37, gen. 2014, doi: 10.1108/JFM-02-2013-0010.

[45] M. Colledani, M. C. Magnanini, e T. Tolio, «Impact of opportunistic maintenance on manufacturing system performance», *CIRP Ann.*, vol. 67, n. 1, pagg. 499–502, 2018, doi: 10.1016/j.cirp.2018.04.078.

[46] Q. Chang, J. Ni, P. Bandyopadhyay, S. Biller, e G. Xiao, «Maintenance Opportunity Planning System», *J. Manuf. Sci. Eng.*, vol. 129, n. 3, pagg. 661–668, giu. 2007, doi: 10.1115/1.2716713.

[47] X. Gu, S. Lee, X. Liang, M. Garcellano, M. Diederichs, e J. Ni, «Hidden maintenance opportunities in discrete and complex production lines», *Expert Syst. Appl. Int. J.*, vol. 40, n. 11, pagg. 4353–4361, set. 2013, doi: 10.1016/j.eswa.2013.01.016.

[48] J. Zou, Q. Chang, Y. Lei, G. Xiao, e J. Arinez, «Stochastic Maintenance Opportunity Windows for Serial Production Line».

[49] C. Richard Cassady, R. O. Bowden, L. Liew, e E. A. Pohl, «Combining preventive maintenance and statistical process control: a preliminary investigation», *IIE Trans.*, vol. 32, n. 6, pagg. 471–478, giu. 2000, doi: 10.1080/07408170008963924.

[50] T. G. Yeung, C. R. Cassady, e K. Schneider, «Simultaneous optimization of $\bar{X}$ control chart and age-based preventive maintenance policies under an economic objective», *IIE Trans. Inst. Ind. Eng.*, vol. 40, n. 2, pagg. 147–159, 2008, doi: 10.1080/07408170701592515.

[51] K. Linderman, K. E. McKone-Sweet, e J. C. Anderson, «An integrated systems approach to process control and maintenance», *Eur. J. Oper. Res.*, vol. 164, n. 2, pagg. 324–340, lug. 2005, doi: 10.1016/j.ejor.2003.11.026.

[52] J. Daaboul, C. Da Cunha, A. Bernard, e F. Laroche, «Design for mass customization: Product variety vs. process variety», *CIRP Ann.*, vol. 60, n. 1, pagg. 169–174, 2011, doi: 10.1016/j.cirp.2011.03.093.

[53] L. Wan e T. Pan, «Double EWMA controller for semiconductor manufacturing processes with time-varying metrology delay», in *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, giu. 2015, pagg. 394–397. doi: 10.1109/CYBER.2015.7287969.

[54] I. Huh, «Multivariate EWMA Control Chart and Application to a Semiconductor Manufacturing Process», pag. 105.

[55] Y. Kuo, «Optimal adaptive control policy for joint machine maintenance and product quality control», *Eur. J. Oper. Res.*, vol. 171, n. 2, pagg. 586–597, giu. 2006, doi: 10.1016/j.ejor.2004.09.022.

[56] S. Panagiotidou e G. Tagaras, «Optimal integrated process control and maintenance under general deterioration», *Reliab. Eng. Syst. Saf.*, vol. 104, pagg. 58–70, ago. 2012, doi: 10.1016/j.ress.2012.03.019.

[57] P. Charongrattanasakul e A. Pongpullponsak, «Minimizing the cost of integrated systems approach to process control and maintenance model by EWMA control chart using genetic algorithm», *Expert Syst. Appl.*, vol. 38, n. 5, pagg. 5178–5186, mag. 2011, doi: 10.1016/j.eswa.2010.10.044.

[58] A. Al-Shayea, M. A. Noman, E. A. Nasr, H. Kaid, A. K. Kamrani, e A. M. El-Tamimi, «New Integrated Model for Maintenance Planning and Quality Control Policy for Multi-Component System Using an EWMA Chart», *IEEE Access*, vol. 7, pagg. 160623–160636, 2019, doi: 10.1109/ACCESS.2019.2950815.

[59] H. Rostami, «Equipment Behavior Modelling for Fault Diagnosis and Deterioration Prognosis in Semiconductor Manufacturing», pag. 162.

[60] S. H. Huang e W. H. VerDuin, «System health monitoring and prognostics - a review of current paradigms and practices», pag. 13.

[61] R. Ahmad e S. Kamaruddin, «An overview of time-based and condition-based maintenance in industrial application», *Comput. Ind. Eng.*, vol. 63, n. 1, pagg. 135–149, ago. 2012, doi: 10.1016/j.cie.2012.02.002.

[62] A. Muller, M.-C. Suhner, e B. Iung, «Formalisation of a new prognosis model for supporting proactive maintenance implementation on industrial system», *Reliab. Eng. Syst. Saf.*, vol. 93, n. 2, pagg. 234–253, feb. 2008, doi: 10.1016/j.ress.2006.12.004.

[63] K.-C. Yang, C. Yang, P.-Y. Chao, e P.-H. Shih, «Applying Artificial Neural Network to Predict Semiconductor Machine Outliers», *Math. Probl. Eng.*, vol. 2013, pag. e210740, nov. 2013, doi: 10.1155/2013/210740.

[64] A. Jalali *et al.*, «Predicting Time-to-Failure of Plasma Etching Equipment using Machine Learning». arXiv, 16 aprile 2019. Consultato: 6 giugno 2022. [Online]. Disponibile su: http://arxiv.org/abs/1904.07686

[65] Tulio Tolio, «Manufacturing system engineering course material held by Tulio Tollio & Maria Chiara Magnanini».

[66] Seleim, A. Azab, e T. AlGeddawy, «Simulation Methods for Changeable Manufacturing», *Procedia CIRP*, vol. 3, pagg. 179–184, 2012, doi: 10.1016/j.procir.2012.07.032.

[67] «Application of descrete event simulation for assembly process optimization».

[68] R. B. Detty e J. C. Yingling, «Quantifying benefits of conversion to lean manufacturing with discrete event simulation: A case study», *Int. J. Prod. Res.*, vol. 38, n. 2, pagg. 429–445, gen. 2000, doi: 10.1080/002075400189509.

[69] M. Jahangirian, T. Eldabi, A. Naseer, L. K. Stergioulas, e T. Young, «Simulation in manufacturing and business: A review», *Eur. J. Oper. Res.*, vol. 203, n. 1, pagg. 1–13, mag. 2010, doi: 10.1016/j.ejor.2009.06.004.

[70] T. Tolio, A. Matta, e S. B. Gershwin, «Analysis of two-machine lines with multiple failure modes», *IIE Trans.*, vol. 34, n. 1, pagg. 51–62, gen. 2002, doi: 10.1080/07408170208928849.

[71] T. A. M. Tolio e D. M. C. Magnanini, «Corso di Laurea Magistrale in Ingeneria Gestionale», pag. 109.

[72] L.-F. Wang e L.-Y. Shi, «Simulation Optimization: A Review on Theory and Applications», *Acta Autom. Sin.*, vol. 39, n. 11, pag. 1957, 2013, doi: 10.3724/SP.J.1004.2013.01957.

[73] N. Rezg, X. Xie, e Y. Mati, «Joint optimization of preventive maintenance and inventory control in a production line using simulation», *Int. J. Prod. Res.*, vol. 42, n. 10, pagg. 2029–2046, mag. 2004, doi: 10.1080/00207540310001638235.

[74] G. Cheng e L. Li, «Joint optimization of production, quality control and maintenance for serial-parallel multistage production systems», *Reliab. Eng. Syst. Saf.*, vol. 204, pag. 107146, dic. 2020, doi: 10.1016/j.ress.2020.107146.

[75] A. A. E. Cadi, A. Gharbi, B. Bouslah, e A. Artiba, «Optimal joint production, maintenance and quality control for a single machine with random failures using Matlab/Simulink optimization- simulation», pag. 8.

[76] L. G. de Oliveira, A. P. de Paiva, P. P. Balestrassi, J. R. Ferreira, S. C. da Costa, e P. H. da Silva Campos, «Response surface methodology for advanced manufacturing technology optimization: theoretical fundamentals, practical guidelines, and survey literature review», *Int. J. Adv. Manuf. Technol.*, vol. 104, n. 5–8, pagg. 1785–1837, ott. 2019, doi: 10.1007/s00170-019-03809-9.

[77] P. Lavoie, A. Gharbi, e J.-P. Kenné, «A comparative study of pull control mechanisms for unreliable homogenous transfer lines», *Int. J. Prod. Econ.*, vol. 124, n. 1, pagg. 241–251, mar. 2010, doi: 10.1016/j.ijpe.2009.11.022.

[78] H. Rivera-Gómez, A. Gharbi, e J. P. Kenné, «Joint production and major maintenance planning policy of a manufacturing system with deteriorating quality», *Int. J. Prod. Econ.*, vol. 146, n. 2, pagg. 575–587, dic. 2013, doi: 10.1016/j.ijpe.2013.08.006.

[79] B. Dassen, A. D. Bucchianico, L. Troisi, e S. Schepens, «Detecting Abnormal Behavior in Lithography Machines», pag. 83.

[80] «Predictive Maintenance: A Reality for Semiconductor Manufacturing», *Critical Manufacturing*, 7 ottobre 2021. https://www.criticalmanufacturing.com/blog/predictive-maintenance-a-reality-for-semiconductor-manufacturing/ (consultato 24 giugno 2022).

[81] A. Nutsch *et al.*, «Characterization of Organic Contamination in Semiconductor Manufacturing Processes», in *AIP Conference Proceedings*, Albany (New York), 2009, pagg. 23–28. doi: 10.1063/1.3251227.

[82] «Optimized Molecular Contamination Monitoring for Lithography», *Particle Measuring Systems*. https://www.pmeasuring.com/it/application-notes/optimized-molecular-contamination-monitoring-for-l/ (consultato 24 giugno 2022).

[83] «Evolving environmental requirements for lithography manufacturing», *Praecis Inc.* https://www.praecis.com/environmental-control-for-lithography (consultato 24 giugno 2022).

[84] «6b869dc2-fde2-48f8-a10d-1e2e7f2cfc9a.pdf». Consultato: 24 giugno 2022. [Online]. Disponibile su: https://www.pmeasuring.com/PMS/files/6b/6b869dc2-fde2-48f8-a10d-1e2e7f2cfc9a.pdf

[85] L. Wang, M. G. Mehrabi, e E. Kannatey-Asibu, «Hidden Markov Model-based Tool Wear Monitoring in Turning», *J. Manuf. Sci. Eng.*, vol. 124, n. 3, pagg. 651–658, ago. 2002, doi: 10.1115/1.1475320.

[86] «6.3.2.4. EWMA Control Charts». https://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm (consultato 13 giugno 2022).

[87] J. M. Lucas e M. S. Saccucci, «Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements», *Technometrics*, vol. 32, n. 1, pagg. 1–12, 1990, doi: 10.2307/1269835.

[88] J. Kinghorst *et al.*, «Hidden Markov model-based predictive maintenance in semiconductor manufacturing: A genetic algorithm approach», in *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*, ago. 2017, pagg. 1260–1267. doi: 10.1109/COASE.2017.8256274.

[89] B. Klaasse, «Condition-based maintenance policies using hidden Markov models», pag. 73.

[90] L. R. Rabiner, «A tutorial on hidden Markov models and selected applications in speech recognition», *Proc. IEEE*, vol. 77, n. 2, pagg. 257–286, feb. 1989, doi: 10.1109/5.18626.

[91] «4.3.1. What is design of experiments (DOE)?» https://www.itl.nist.gov/div898/handbook/pmd/section3/pmd31.htm (consultato 22 giugno 2022).

[92] M. Uy e J. K. Telford, «Optimization by Design of Experiment techniques», in *2009 IEEE Aerospace conference*, mar. 2009, pagg. 1–10. doi: 10.1109/AERO.2009.4839625.

[93] «Application_of_Design_of_Experiments_to_Plasma_Arc.pdf».

List of Figures

List if Tables

List of Acronyms

APC Advanced process control

SVID status variable identification

PCA principal component analysis

SPC statistical process control

OOC out of control

RUL Remaining useful life

DES Discrete event simulation

2M1B Two machine one Buffer

IC integrated circuit

AMC Airborne molecular contamination

MTTF Mean time to failure

MTTR Mean time to repair

IoT internet of things