POLITECNICO DI MILANO
DEPARTMENT OF MECHANICAL ENGINEERING
MASTER THESIS

An Analytical Model to Determine Reaction Levels in a Lean Production System

Master Thesis of
Paolo PAGANI

Supervisor :
Tullio TOLIO

Abstract

In recent decades the lean production systems have been developed and studied worldwide due to their success in the Toyota company and due to the well known advantages that those systems can bring to the company, such as to have low costs associated to the WIP but at the same time a high level of efficiency and customer service level.

Lean manufacturing systems are usually designed and sized for a predetermined production volume or customer demand respectively. In the design phase and during the periodical new parameters choice, some assumptions about process characteristics in the future must be made, i.e. transportation times, OEE-levels, setup times, demand fluctuation, etc. During the operation phase it can happen that some assumptions are not verified or change during the whole considered period in the plant. Usually those deviations are on the worse side but possibly also on the better side. In order to keep the system lean, i.e. with a rigorous attention on all costs (also the hidden ones which are often not considered) and on the quantities of produced parts but keeping a high customer service level, those lean production systems call for escalation levels. It means that some variables must be monitored, their safe ranges must be defined and when their actual values exit the normal range, some effective reaction policies must be implemented to bring back the values in the safe ranges. Classical lean production systems have been designed with fixed escalation levels but since the optimal levels are strongly dependent on the productive environment, it would be better to update them as soon as some characteristics of the production system or of the demand change. For that reason some fast and efficient tools are required to compute them when required. In this master thesis an already existing analytical model based on mixed state Markov Chains is suggested to study Kanban-based production systems and it is explained how the model

can fit typical industrial situations. It is practical and fast to be used because it returns an exact and immediate solution by requiring only a small amount of data as input. As a consequence, it can be applied to a large variety of real situations. The model is created by fitting that model on the real case of a Bosch production plant located in Bari (Italy), in which the BPS (Bosch production system) characteristics along with the related escalation policy, based on the number of already finished products in the final warehouse, are implemented.

The master thesis is divided in 4 parts:

First of all, the lean production systems are introduced and it is explained what are the typical critical points in the modern competitive context. The uncertainty types and the most typical risk scenarios that influence their performances and the decision levels that can help to tackle and limit those problems are then described. In addition, a general guideline for the design of a robust system is presented.

Secondly, it is defined which subset of problems will be investigated in this work and why they represent some of the crucial problems in the modern lean production systems. Once the aim of this work is defined, the exact analytical method which will be used to tackle the problems is recalled and described. In particular, it is explained the procedure to model the defined system, to obtain its performances starting from the input parameters and how it can be helpful in the computation of its optimal reaction levels by defining and minimizing a cost function. The method models a system composed by two machines, which are decoupled by a buffer with finite capacity. The upstream machine represents the production stage, which stores parts in the final warehouse and can be reconfigured according to the current buffer level when the inventory level is either too low or too big, and the downstream machine represents the withdrawal behavior of the customer.

Thirdly, the Bosch case is presented and modeled. It is also shown how the cost evaluation for the current reaction policy can be performed and the expected improvements if the optimal reaction levels would be computed with the just developed tool.

In the last part, it is shown what would happen if the production environment, the demand or cost parameters had different characteristics and, in particular, how the optimal threshold would change and how much extra costs would the company incur if the reaction policy would not be adapted to the new production context characteristics.

Acknowledgements

The period spent writing my Master Thesis has been exciting and full of new experiences for the professional and personal point of view. For this reason there are several people to whom i want to express my deepest acknowledgment for the help they provided me, during the course of my work.

First and foremost, i would like to thank my advisor Prof. Tullio Tolio for his patient guidance, enthusiastic encouragement and useful critiques. His wide experience in the manufacturing problems taught me how to rigorously face and tackle the real production problems. I am also grateful for the opportunity that he gave me to write my thesis in collaboration with the Karlsruher Institut für Technolgie (KIT) in Germany. The international experience gave me the possibility to become proficient in a new foreign language and to learn the way of thinking of another country.

I would like to express my very great appreciation to Prof. Dr. Kai Furmans, head of the Institute of Material Handling Systems and Logistics in KIT, for having hosted me in his institute for 9 months, for having shared with me important knowledge and information and for having made me write my Thesis in collaboration with the Bosch company. I receive a precious support also from his assistant Ing. Martin Epp, who has helped me during the whole period spent in Germany.

Next, i wish to thank the people of the Bosch production plant located in Feuerbach. They cooperated in my project, made me visit the production plants and shared with me important information about the Bosch production policies. Among them, i especially express my gratitude to Ing. Marco Morea for his constant availability and assistance.

The advice given by Ing. Andrea Ratti has been a great help in the development of my thesis, especially during the time periods spent in Italy. He gave me important

suggestions and guidelines in some crucial parts of my Thesis.

A special thank goes to my parents Ernesto and Vittoria, my sister Barbara, my brother-in-law Elia. They assisted me during the important choices of my life, encouraged me during the whole course of study, allowed me to do this international experience and, last but not least, they set a good example to me.

Last but not least, a huge thank to all my friends, the ones who shared with me the experience in Karlsruhe, the ones who spent with me the university years and the ones who shared with me the important moments of everyday life.

An Analytical Model to Determine Reaction Levels in a Lean Production System

Paolo Pagani

September 12, 2014

Contents

1 Introduction **1**
- 1.1 General overview on lean production systems . . . 1
- 1.2 Typical problems of a lean production system . . . 2
- 1.3 Reaction and escalation policies . . . 3
- 1.4 Reaction policy design questions . . . 4
  - 1.4.1 Question 1 . . . 4
  - 1.4.2 Question 2 . . . 5
  - 1.4.3 Question 3 . . . 6
    - 1.4.3.1 Production configurations and range combinations . . . 6
    - 1.4.3.2 Decisional delays . . . 8
    - 1.4.3.3 Optimal reaction policy . . . 9
- 1.5 Cost optimization . . . 11

2 Motivation and problem statement **13**
- 2.1 Motivation . . . 13
- 2.2 Problem statement . . . 14
- 2.3 Goals and objectives . . . 15
- 2.4 Overview of the study . . . 15

3 Literature review **17**
- 3.1 Supply chain risk management . . . 17
- 3.2 Development of lean production . . . 18
- 3.3 Adaptive and reactive lean production systems . . . 21
- 3.4 Concept of thresholds . . . 23
- 3.5 Modeling of the production rate distribution . . . 25

4 Methodology **27**
- 4.1 Description of the system to be modeled . . . 27
- 4.2 Multi-threshold model . . . 28
  - 4.2.1 Conceptual model description . . . 28
    - 4.2.1.1 Assumption and description . . . 28
    - 4.2.1.2 Inputs and output . . . 29
- 4.3 Modeling of the considered system . . . 29
  - 4.3.1 Threshold number and negative inventory level . . . 30
  - 4.3.2 Upstream machine - production system modeling . . . 31

4.3.3 Downstream machine - demand modeling . . . 34
4.3.3.1 Demand time series . . . 34
4.3.3.2 Scenarios with multiple states . . . 35
4.3.4 upper and lower threshold transitions . . . 35
4.3.5 Tools for behavior modeling with Markov chains . . . 37
4.3.5.1 PH Fit tool . . . 37
4.3.5.2 Modeling with only 1 up and 1 down state . . . 37
4.4 Model outputs . . . 40
4.5 Objective function . . . 41
4.5.1 Optimization algorithm . . . 43
4.6 Use of the model in the decision-making . . . 44

5 Bosch real case **47**
5.1 Introduction . . . 47
5.2 Products . . . 47
5.3 Plant layout description . . . 49
5.4 Customer . . . 52
5.4.1 Available demand information . . . 52
5.4.2 Stockout and backlog . . . 53
5.5 Kanban policy . . . 53
5.5.1 Computation of the Kanban number . . . 55
5.5.2 Considerations on the Kanban computation . . . 58
5.5.3 Usage of the Kanban policy . . . 59
5.6 Production policy . . . 60
5.6.1 Available time to produce computation . . . 61
5.6.2 Reaction policy . . . 62
5.6.2.1 Escalation ranges . . . 62
5.6.2.2 Adjustable parameters - number of shifts . . . 63
5.6.2.3 Daily schedule and leveling . . . 64
5.6.2.4 Considerations on the current reaction policy . . . 66
5.7 Anticipated replanning . . . 66

6 Bosch case modeling **69**
6.1 Upstream machine modeling . . . 69
6.1.1 Available data . . . 69
6.1.2 Rate distribution . . . 72
6.1.2.1 non-parametric ANOVA on the number of daily shifts . . . 74
6.1.2.2 non-parametric Anova on the week days . . . 76
6.1.3 Production plan characterization . . . 80
6.1.4 Decision delay modeling . . . 83
6.2 Escalation levels . . . 83
6.3 Downstream machine modeling . . . 84
6.3.1 Considered demand profile . . . 85
6.3.2 Modeling of the scenarios . . . 86

6.3.3 Modeling of the scenario transitions . . . 89
6.4 Cost evaluation and optimization . . . 90
6.4.1 Cost coefficient . . . 90
6.4.2 Optimization . . . 93
6.5 Biweekly decision delay . . . 95
6.5.1 Delay length comparison . . . 96

7 Optimal thresholds considerations 101
7.1 Cost coefficients . . . 101
7.1.1 Inventory costs . . . 102
7.1.2 Backlog costs . . . 103
7.1.3 Costs of the production plans . . . 105
7.1.3.1 Cost of manpower . . . 105
7.1.3.2 Overtime costs . . . 106
7.1.3.3 Manpower flexibility . . . 107
7.2 Production plan characteristics . . . 109
7.2.1 Production expected value . . . 109
7.2.2 Production variance . . . 111
7.3 Demand characteristics . . . 112
7.3.1 Demand expected value . . . 112
7.3.2 Daily demand regularity . . . 114
7.3.3 Demand extra-variance . . . 115
7.4 Maximal inventory level . . . 117
7.5 Number of reaction policies . . . 118
7.6 Number of reaction policies - simplified model with biweekly decision . . 121

8 Conclusions and future research 123
8.1 Work summary . . . 123
8.2 Research contributions . . . 124
8.3 Considerations for the company . . . 125
8.4 Future developments . . . 127

1 Introduction

1.1 General overview on lean production systems

Traditionally lean production systems are designed to be cost-efficient by implementing some management tools and policies, which aim to reach goals that in a normal productive context would be impossible to be obtained at the same time and by achieving a radical cultural change in the way all people in the company think. Those goals are usually the elimination of the inefficiencies in the productive context, high quality, high flexibility, high productivity, low lead time and low production costs. Those goals seem obviously conflicting and that is the reason why they can be reached only with effective control policies and a precise way of thinking of both workers and managers. A lean production system must be designed from the beginning, starting from the product design phase along with the production plant design. In fact, the simplicity and the standardization of the components which compose a product along with a reliable production plan which allows quick setups are key factors to achieve those contrasting goals. All those goals require a great effort to the company but they can significantly improve its competitiveness ([BTS92]).

Moreover, proper decisions for what concerns the supply chain, which must include suppliers who deliver the components and the raw material within the scheduled time and as flexible as possible, must be taken. Then the last operative phase must be also well designed. Some different organization policies must be implemented, such as a wide use of the so called "job enlargement" and "job enrichment". It means that on one hand the workers must be trained to be flexible and able to make different tasks when it is needed. On the other hand they must be able to evaluate their own work and to take some basic operative decisions without necessarily asking the production managers. That makes the intervention faster and the problem is solved sooner. Furthermore, the manpower must be ready to work overtime or to work less when required, for example, when the demand has some weekly fluctuations which would increase the risk of having stockout or high inventory level, some extra shifts or shift removals can be planned in advance to follow the trend of the demand. Finally, the operational management must implement some useful techniques to plan, modify and control the production. Some examples are the leveling of the scheduled production, it means that, if the control on the weekly production volumes by adding and removing shifts (or other control policies) is not enough or is less economical, the management can try to make the system produce at a more constant rate by scheduling the production of a certain amount of parts on days where a lower quantity is required by the costumer and by making the final warehouse absorb the demand fluctuations.

Another example of control policy on the production volumes are the Kanban-based production systems. They consist in attaching some cards to each product that is taken from the upstream supermarket of a certain production area (for example, a department), then the product is stocked into the downstream supermarket and when it is withdrawn the card is removed and sent back to the beginning of the area ready to be assigned to a new product. Since the number of card is fixed and the workers are not allowed to produce more if no free cards are available, this method limits the value of the work in progress in that area, forces the production rate to automatically change when some demand changes occur and gives an immediate feedback to the production planner about which products risk the stockout and which ones are not needed to be produced.

1.2 Typical problems of a lean production system

Generally, all those strategies work well in a repetitive environment. However, the production system can hardly remain at a high performance level when the production context becomes very unstable. Modern supply chains must face a variety of risks and they often operate in an uncertain environment. Especially in recent years, since the global competition has increased and the customers require more customizable products and shorter delivery times, companies are forced to improve their supply chains by implementing some policies, which can tackle better risks. In particular there is a large variety of risks and they can generally be divided in two categories. The first one includes some small and frequent problems like normal demand fluctuation, process parameter (OEE, MTTF, MTTR, transportation times) variance, etc. They affect the system performances in a limited way [Tan06a] because the system is generally already designed to tackle small unexpected environment changes. For example, the production rate can be adjusted to follow a demand with small fluctuation. By contrast, the second one, which includes some uncommon but serious production environment changes, can cause several problems and the performances of the system can dangerously get worse. The production systems, especially the lean ones, are usually not designed to tackle those kinds of problems because they are not taken into consideration in the design phase, since they are generally quite rare and during the plant life-cycle the company will just try to make them as small a possible. For example, the company will choose the suppliers not only considering the cost of the supply but considering also its reliability. In any case, those risks can not be completely deleted and when they occur, the problems for the company are huge. In this group of risks we can include unexpected mean value variation of the customer demand (for example, due to an economical crisis or to the variation of the costumer number or the costumer weekly required volumes), economic turbulence, serious production problems due to the increase of the failure rate of some machines, etc. Although those risks occur with a low probability, their impact can be catastrophic. The requirement of implementing efficient policies, which react and adjust the system when some risk scenario occur, has become important and inevitable. For instance, [Lee04] has studied more than 60 leading companies and suggested that the design of a triple-A (agility, adaptability and alignment) supply chain was an important

factor in determining the success of a company in a modern uncertain environment.

1.3 Reaction and escalation policies

In order to build a robust supply chain network and limit the impact of risks, a series of robust approaches and policies have been developed and optimized by researches and according to the current literature those approaches can be divided in 4 categories.

First of all, supply management, which includes the proper choice of one or more suppliers considering both their costs associated with the purchasing of raw material and components and the ones related to their reliability. It means that a supplier can provide components at a very competitive price but still not be a good choice because he usually delivers with a delay or the wrong quantity because he is far away and the long transportation times add variability and uncertainty to the deliver date. Secondly, demand management, which includes the product dynamic pricing to make the demand as stable as possible, i.e. when the demand decreases, the possibility of decreasing the final product price and vice versa is taken into consideration and its convenience evaluated. Thirdly, production management, which includes the production leveling. It means that if the demand is not regular, some production volumes can be shifted in order to have a more regular daily production. Lastly, information management, which includes the demand collaborative forecasting, namely the sharing of information through all the entities of the supply chain to make the forecast more accurate.

Moreover, the decision making activities can be grouped into three levels : strategic, tactical and operational([Bal92] ; [SW00]). The first category regards all the possible decisions that affect the structure of the supply chain, for example, the choice of the suppliers or of the production site. For that reason, once they are taken, they freeze some supply chain characteristics in the long-term (usually more than one year) because, in order to create an efficient and effective one, a stable supply chain structure and a strong relationship between a company and its suppliers or costumers must exist. The second category is represented by the tactical ones, which are related to the mid-term (usually one or more months) and can include the production plan schedule for the next month, allocate the workload for service facilities and assign transportation routes for distribution centers and so on. At last, there is the operational level, which refers to the short-term (usually less than 2 weeks) decisions like the schedule of the weekly production plan and transport activities, the forecast of the weekly customer demand or the continuous monitoring of inventory level.

The amount of literature dealing with robust models to design and manage supply chains in an uncertain environment is quite big but most of those models focuses only on the strategic and tactical level, such as to determine the facility locations or the plant capacity ([Lim09]). Only a few models deal with the operational level, such as the modified Kanban system models, which can flexibly change its number of Kanban cards in response to the customer demand fluctuations ([TN99] ; [Tak03]). In any case, those operational level models always consider simple and typical risk scenarios and neither severe demand fluctuations nor scenarios related with supplier or process uncertainties

risks (e.g. supplier material shortage, process machine breakdown).

An exception is [Li13], that presents an empirical approach to model Kanban Systems with escalation levels, namely that some indexes, which are related to the demand rate with respect to the production rate of the system bottleneck and to the inventory level of finished products, are defined and their current smoothed values are monitored. When they reach dangerous values, for example when the demand becomes too big with respect to the bottleneck production rate or when the inventory level becomes too low, it is economically estimated, if and how it is economical to activate some escalation reactions. This model includes both the supply management and the product management activities, the reaction policies include all the three levels (strategic, tactical and operation) and it is considered that the more dangerous the actual situation is, the stronger the reaction that is taken into consideration is. For example, if the inventory level is a little lower than the safe range an operational reaction will be considered but if it is dangerously low a strategic one will be considered instead because, although it costs more, it avoids some bigger costs like stockout and backlog costs, which they are more likely in case of dangerously low inventory level.

1.4 Reaction policy design questions

In [Li13] the topic of how a robust and lean production system can be designed is also introduced and it is stated that 3 basic questions must be answered:

- Which parameters and characteristics of the production system must be dynamically changed to improve the performances and, in general, to maximize the net profit?
- Which indexes must be continuously monitored or constantly evaluated to find out if some reaction policies must be implemented?
- When and how to react depending on the actual value of all the monitored indexes?

1.4.1 Question 1

The answers to all the three questions depends strongly on which production system it is considered. Nevertheless, a set of typical examples can be provided.

Typical system parameters which can be dynamically changed are:

1. supplier aspects:
   - main supplier
   - backup supplier
2. demand aspects:
   - dynamical pricing

3. amount of information sharing along the supply chain:
    - inventory levels
    - production status
    - accuracy of delivery dates
    - accuracy of the customer required parts
4. production characteristics:
    - machine nominal production rates
    - active machines' number
    - available manpower, overtimes and shifts number
    - number of production and transportation Kanbans
    - maintenance plan

It is important to notice that some of them are easy and fast to be implemented such as the variation of the number of Kanbans, while others take longer times and greater economical efforts. As already said, they can be divided in three categories : operational-level, tactical-level and strategic-level adjustable parameters. The policies which involve the first ones are fast and cheap ways to respond to slight risks but they are not so effective when the system is already far from safe operation conditions. On the other hand, the ones involving the second class of parameters (and even more with the third class) takes a longer time and cost more but they can avoid greater costs when the risk is bigger. However, it is generally suggested (as shown in [Li13]) to design a reaction policy which involves all the three kinds of activities for different risk levels.

1.4.2 Question 2

In order to answer the second question, we must define what are the goals of the production plant, which are the risks that the company wants to avoid and what indexes can properly detect them. In a lean system the typical goals are:

- high service level
- low production costs
- low inventory levels and work in progress
- low scraps
- low lead time
- high quality

If the company wants to achieve them, the typical indexes which must be monitored are:

- inventory level, which must be kept, on one hand, not to low to limit the stockout risk and the backlog costs, which causes a low service level, and, on the other hand, not too high because it implies high inventory costs

- ratio between the demand rate and system bottleneck rate. In fact, if it is greater than one and it is a systematic problem, a stockout situation will soon occur. On the contrary, if it is too low it means that the inventory will be soon full and, as a result, the production plant will be blocked because it can not store the produced parts and the inventory costs will increase

- percentage of good parts, which detects, if there are some quality problems in the production plant

- mean delay of the main supplier, which could detect if there some problems with the deliveries of raw material and components

From a theoretical point of view, the number of controlled parameters, which can bring some benefits is very big for a complex system like a production plant. However, since the monitoring process and the data analysis imply additional costs, the only parameters which is wise to consider are ones whose economical benefits are bigger than the monitoring costs. Moreover, as it can be seen in the literature, during the design of a robust reaction policy it is always preferable to deal with one or two parameters in order to keep the model simple and clear.

1.4.3 Question 3

1.4.3.1 Production configurations and range combinations

By considering all the above-listed reconfiguration parameters, it must be now decided which ones will be considered in the building process of a robust system. Let us say that M reconfiguration parameters, defined by the symbol P_m, are chosen and for each of them l_m possible levels are defined. As a result, $C = \prod_{m=1}^{M} l_m$ different production configurations, i.e. all the possible combinations of the defined levels of the chosen parameters, can be defined. Then, a set of N monitored indexes (denoted by the label I_n) can be defined and for each I_n index the whole possible range is divided in R_n smaller ranges with $r_n \in (1, 2, .., R_n)$. For each range combination $(r_1, r_2, ..., r_N)$ one and only one preferable configuration is associated. The escalation levels are the boundaries which define all the possible range combinations. Let us provide an simple example where there are two monitored indexes $I_1 \in [0, 1]$ and $I_2 \in [0, \infty)$ and $R_1 = 2$ and $R_2 = 3$. The range separation can be, for instance, as follows:

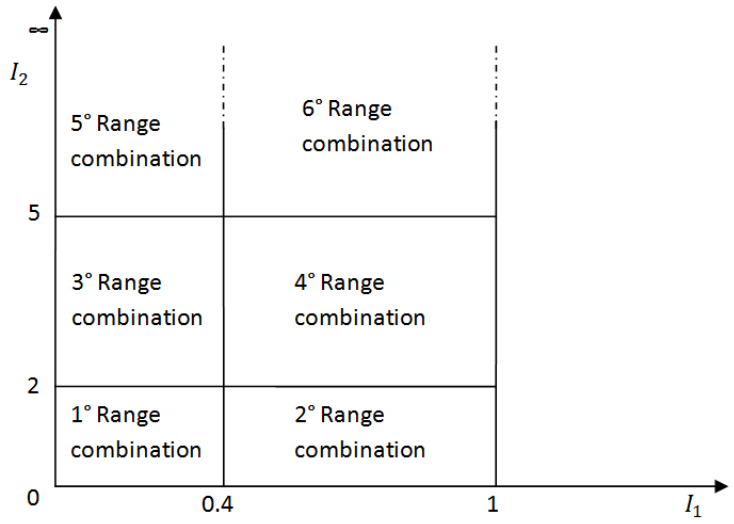

Figure 1.1: range combination regions for a 2 indexes example

In the example 6 regions in the space defined by the two indexes are created and for each region a single preferable production configuration must be assigned. In this case let us suppose that there are 4 possible different configurations ($C = 4$) and they are assigned as follows (it is not compulsory to assign all the possible configurations and, in particular the subset of used configurations is denoted as C' which in this can coincides with C) :

$Numberproductionconfiguration$	$conditiononI_1$	$conditiononI_2$
$1^{st}productionconfiguration$	-	$0 \leq I_2 \leq 2$
$2^{nd}productionconfiguration$	-	$I_2 \geq 5$
$3^{rd}productionconfiguration$	$0 \leq I_1 \leq 0.4$	$2 \leq I_2 \leq 5$
$4^{th}productionconfiguration$	$0.4 \leq I_1 \leq 1$	$2 \leq I_2 \leq 5$

It means that the correspondence between range combinations and production configurations is as follows:

Number range combination	r_1	r_2	range r_1	range r_2	preferable configuration
1^{st} Range combination	1	1	0 - 0.4	0 - 2	1
2^{nd} Range combination	2	1	0.4 - 1	0 - 2	1
3^{rd} Range combination	1	2	0 - 0.4	2 - 5	3
4^{th} Range combination	2	2	0.4 - 1	2 - 5	4
5^{th} Range combination	1	3	0 - 0.4	5 - ∞	2
6^{th} Range combination	2	3	0.4 - 1	5 - ∞	2

Up to now it has not been considered that, when a threshold is reached, it is not necessary crossed but the monitored indexes can also remain on the threshold. For that

reason, if that phenomenon is possible, the preferable policies on each boundary which delimits each range combination must be defined. Let us provide an example for the considered 2 indexes case. The figure 1.1 becomes as follows:

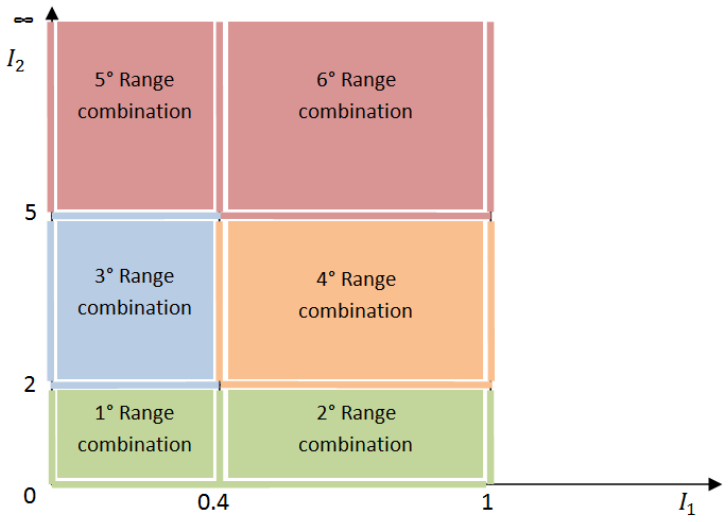

Figure 1.2: range combination regions for the 2 indexes example with preferable configurations

Number of preferable configuration	color
1	green
2	red
3	blue
4	orange

For instance, if the state of the system identified by the current index values is in the 4^{th} range combination (as a consequence the preferable policy will be the 4^{th} one) and P_2 increases until the threshold which delimits it in the upper side is reached, the system can remain on the threshold and in this case, although it is not already in the 6^{rd} range combination, the preferable policy becomes the one related to that threshold, in particular the 2^{nd} one.

1.4.3.2 Decisional delays

Up to now an important aspect has not been considered, namely that in a real production plant, when it is reported that some changes in the production configuration must be carried out, they are not instantaneous implemented but there are some delays. For instance, if the monitored index is the buffer level of final products, it can happen that it decreases until it crosses an escalation level and exits the safe range. As a result, it becomes preferable a policy which implies to activate some more machines to increase the

nominal upstream production rate, make the inventory level return in the normal range and, as a result, to decrease the stockout probability. That decision is not instantaneous, i.e. simultaneous with the threshold crossing or with the arrival on the threshold. In practice, it has generally a certain delay because, for instance, the alarm situation must be firstly detected, then the managers must be informed, they must approve the decision and finally the decision must be implemented, namely the additional machines must be switched on, warmed up and some workers must be assigned to them. Anyway, the delay times for each action can be strongly different depending on many factors (how the threshold crossing is reported, how the approval process is conducted, how long does it take to implement the decision and so on).

In order to make a rigorous reasoning it can be said that, if the system is working with the i^{th} configuration at the time t and immediately after this moment the j^{th} configuration becomes preferable (for instance, due to a threshold crossing), the change will be implemented with certain delay distribution. It can be a deterministic delay value, a continuous or a discrete random variable.

In general a delay matrix [NxN], which contains all the delay distributions to switch from the current i^{th} to the j^{th} production configuration, can be defined as follows:

$$delay_{i,j} = \begin{pmatrix} 0 & \hat{\mathrm{d}}_{1,2} & \cdots & \hat{\mathrm{d}}_{1,N} \\ \hat{\mathrm{d}}_{2,1} & 0 & \cdots & \hat{\mathrm{d}}_{2,N} \\ \vdots & \vdots & \ddots & \\ \hat{\mathrm{d}}_{N,1} & \hat{\mathrm{d}}_{N,2} & & 0 \end{pmatrix} \tag{1.1}$$

Of course it does not make any sense to represent the delay from and to the same configuration because it implies no changes and for that reason the values on the diagonal are set to zero. It must be noticed that, if the delay distribution is not dependent on the starting configuration, the values of each column are all equal.

Furthermore, it can also happen that the j^{th} configuration remains preferable for a short period of time (or the the reconfiguration takes a long time) that the reconfiguration does not occur and another k^{th} configuration becomes preferable due to another threshold crossing. In order to evaluate the new delay distribution, we just have to consider the random variable in the position (i, k) in the delay matrix without considering what happened before.

1.4.3.3 Optimal reaction policy

Since the company is interested in designing a robust reaction policy which minimize the costs in a given risk scenario and for the assigned production system, the final goal of this reasoning is to find the optimal reaction policy. The problem can be divided in three main tasks which must be carried out simultaneously because they depend on each other:

1. an optimal set of monitored parameter $(I_1, .., I_n)$ along with their optimal range numbers $(R_1, .., R_n)$ and optimal reaction levels must be found

2. an optimal set of adjustable parameters $(P_1, .., P_m)$ along with their optimal number of considered levels $(l_1, .., l_m)$ and the parameter value for each of them must be found

3. the available configurations for the standalone parameters must be combined to obtain all the possible production configurations (C) and the optimal configurations which will be used (C') must be select and properly assigned to each range combination and threshold

When more than one monitored parameter and many, or even infinite, possible system configurations are chosen, it becomes quite hard to find the optimal reaction policy. As a consequence, it is generally chosen to simplify the problem either by providing just a reasonable solution and not the optimal one ([Li13]) or by considering a very small number of adjustable parameters and monitored indexes ([GGF+09] , [GGF+13]). In the first case, some characteristics of the reaction policy are computed first, i.e. without the simultaneous optimization, and the few remaining characteristics are then optimized. For instance, it is possible to define the entire reaction policy except for the threshold level which are then optimized.

The last task, i.e. the assignment of the configurations to the index ranges, is particularly crucial because, if it is not done properly the reaction policy can be ineffective or even make the situation worse. For example, if the inventory level of a warehouse is monitored and three levels of this index are considered and it is wrongly imposed that, when the inventory level is in the low range, the production rate must be decreased and, when it is in the high one, the production rate must be increased, we will optimize a reaction policy which can only make the costs increase.

In any case some general considerations can be done about how the possible configurations can be associated with the index range combinations. However, they must be considered as guidelines which are valid only for the most of the production systems:

- There generally exists for each index one or more desirable ranges and it is preferable to make the system work in those operation conditions because, for example, that implies the lowest operational cost, the lowest unitary cost per finished product or the highest throughput depending on which objective function in considered

- For each range combination a reasonable configuration must be chosen to try to make the system operate in the desirable range again. For instance, it means that, if the costumer is requiring more products and the inventory level, which is the monitored parameter, becomes dangerously low, it is reasonable to increase the production rate as soon as possible

- If two possible configurations have the same implementation delay and effect on the system and their costs are different, the more expensive choice will not be chosen

- If two possible configurations have the same implementation costs and effect on the system but their implementation delay is different, the slower choice will not be chosen

- The impact of a certain decision must be taken into account. In general it can be stated that the more the indexes are far from the "safe ranges", the stronger the reaction must be but usually a configuration which implies greater effects for the system is also more expensive and takes a longer period to be implemented and it must be also taken into account during the reaction policy design

1.5 Cost optimization

It is really important to understand that, when the optimal escalation levels of some monitored parameters must be determined, a proper cost function, which must contain all the differential costs involved in the decision, should be defined and optimized. An optimum for the escalation parameters in the region delimited by the constraints always exists and it generally holds that, on one hand, it is not economical to have a too reactive system, i.e. a system that changes its characteristics as soon as small variations in the production environment occur, if there is a cost associated to the reconfiguration. On other hand, it also not cost-efficient to design a system which never reacts because it would be equal to the situation where no reaction policy is implemented. It is difficult to generalize this statement because everything depends on how big the escalation costs are but in the literature there are many examples where it is shown that this statement holds for typical real cases ([GGF+09] , [GGF+13] , [Li13] and so on). For instance, if the reaction and escalation costs are zero and the reaction policy is reasonable (for example, increase the upstream production rate in some way, when the final inventory becomes dangerously low) and instantaneous, the optimization of the escalation levels suggests a system as reactive as possible. On the other hand, as the reaction costs tend to infinite, it will be better to avoid any escalation policy.

2 Motivation and problem statement

This chapter is divided in four main parts. The first section clarifies the reason why the topic of this work has been chosen and must be further investigated. Secondly, the problem to solve is exactly identified and, thirdly, the goals and objectives which must be achieved with this work are defined. In the final part, a short overview of how the master thesis is structured is provided.

2.1 Motivation

In the last decades the competitive environment has become more and more turbulent because the customers have higher requirements in terms of customization, quality and service and the globalization has made the number of competitors and markets bigger. For example, new products are continuously designed and produced and it implies a great impact on the supply chain performances if no changes are implemented ([PASB10]). As a result, the goal of remaining competitive has become a hard task especially for those companies which consider the lean production as a key factor for their success. As already said, those kinds of production plant are designed and managed to be as cost-efficient as possible and, at the same time, to achieve small lead times, low work in progress, high quality, high flexibility, high productivity, high service level and so on. In order to obtain all those goals at the same time in the competitive environment of nowadays, the traditional lean production systems introduced in [Ohn88] and [Mon83] are not sufficient and some new powerful management tools and policies are required. One of the most used and studied method to achieve the listed goals is to design a reactive system which is able to face some impending risks by modifying some of its characteristics during the operation time. In particular, the most common and relevant risk in manufacturing companies with a make-to-stock policy is to incur the product backlog, which occurs when a customer asks for a certain quantity of products but the demand can not be immediately satisfied. As a consequence, in the literature there are many researches which dealt with this problem by implementing a reaction policy which triggers the upstream production to limit this risk. It means that in those models the inventory level of final products (or an index related to it) is usually monitored to detect that risk and the behavior of the upstream part is changed depending on the current amount of finished parts ([Li13] , [TM01] , [Tak03]). However, the backlog is not the only concern of the company. Indeed, during the whole plant life the goal is to minimize the overall costs which include also the costs related to the inventory, the manpower, the maintenance, the reconfigurations, etc.

However, it has been noticed that in the current literature there not exists a model

which includes all the following characteristics at the same time:

1. Possibility to model a system which is capable to react with operative, tactical and strategic activities and with different reaction delay distributions
2. The reaction policy is based on the monitoring of the inventory level of finished products (by means of thresholds) because the aim is to limit the product stockout and backlog and at the same time to produce the parts as efficiently as possible
3. Possibility to assess the impact of both normal uncertainty sources and disruptive risks
4. The model must require only a limited amount of information as an input
5. Rigorous cost evaluation. In particular, with a great attention to backlog costs, inventory costs and production costs
6. Possibility to optimize the reaction policy and, as a result, to find the best solution for the given problem and not just a good one
7. The model is analytical. It implies that the computation time must small compared to the complexity of the considered system

2.2 Problem statement

The state of art lacks of a model which is capable of finding the optimal reaction policy to design a robust system in a given risk scenario and includes all the aforementioned characteristics.

First of all, the model must be analytical, since it can be generally seen that, although the analytical tools are not able to model systems in details, they generally return an exact solution in a smaller time (especially during the simultaneous optimization of many parameters) than the simulation models and without the problem of handling confidence intervals.

Secondly, it must be able to model not only the normal uncertainties, about which the current literature already contains plenty of researches, but also to evaluate the impact of disruptive events. Many uncertainty sources can also be considered at the same time, such as the ones related to the suppliers, the production plant, the manpower and the demand.

Thirdly, it must require a small amount of information in input because in the real production systems not all the information about failures, maintenance, supplier service and customer behavior are measured. As a result, when some system analysis are required to investigate some properties, it is generally difficult to investigate particular system features. Moreover, even if some data are available, there must be enough to make a relevant inference and it must be assured that the historical data are representative of the current system. To sum up, in practice only few and general pieces of information

are available to estimate the system behavior. For instance, as far as i am concerned, it is generally easy to provide a good estimate of the daily aggregated production, i.e. considering the daily production volumes of an equivalent finished product, but it is quite hard to obtain the same estimate for each different products. Furthermore, it is generally easy to estimate the mean daily downtime of a machine but it is difficult to estimate properly its mean time to failure and to repair.

Fourthly, the most important index to be monitored must be the inventory level of finished products, since the main goal of a company is usually to limit the backlog and, at the same time, the inventory level. The company must achieve that by minimizing also the production costs, which include manpower, maintenance, reconfiguration, material and quality costs. All the costs which depend on the chosen reaction policy must be included in the cost function for the evaluation and optimization.

Lastly, it must be possible to define control limits which identify ranges for the monitored index and to associate a different preferable system behavior for each of them. For example, if the buffer is almost empty, it is wanted to make the upstream system produce more and vice versa. In particular, if the monitored index crosses a control limit, it means that the preferable configuration has changed and the system will not reconfigure immediately but with a certain delay distribution. It must be also possible to define different delay distributions depending on both the starting and arrival configurations.

2.3 Goals and objectives

This thesis aims to provide general guidelines for the modeling of robust production systems, namely the typical questions that should be answered during the rigorous procedure to find a useful reaction policy and their usual solutions for the manufacturing context. After that, an analytical tool, based on the model presented in [TR13], will be proposed to deal with the problem defined in the previous section because it can potentially meet all the requirements. Then, the Bosch real case must be introduced and it must be shown how the general reasoning can be applied to a real case, which provides real problems and real data. In particular, it must be proved that the model is useful and can help the company to be more cost-efficient. Lastly, it must be shown how the thresholds change and how much money can be saved changing them when the competitive context varies to demonstrate the convenience of adapting the control policy to it.

2.4 Overview of the study

The steps of this work can be divided in two main parts: a general methodology to deal with the chosen subset of problems and its application on Bosch real case.

In the first part the topics are organized as follows:

1. Preliminary questions for the design phase of a reaction policy

2. Modeling of the different system configurations
3. Modeling of the reaction delays
4. Modeling of the customer behavior
5. Procedure to obtain the different system performances for a defined reaction policy starting from the model outputs
6. Cost evaluation considering the system performances
7. Cost optimization

In the second part the Bosch case is introduced and studied:

1. Description of the products, layout and customer behavior
2. Current Kanban and reaction policy
3. Modeling of the described system
4. Cost evaluation with the current production policy
5. Cost optimization and improvement assessment
6. Threshold trends as a function of the system parameters and extra costs if the reaction policy is not changed

3 Literature review

In this chapter the main contributions related the topics of supply chain risk management, lean production and reactive lean production systems and production distribution modeling are recalled and presented. The purpose is to provide a preliminary knowledge about what has already been studied and what is still missing.

3.1 Supply chain risk management

The risk management along the whole supply chain of a company has becomes in the last decades a central issue. It happened mainly because competitive context has been getting more and more turbulent and complex. For example, the global competition, the customer requirements, the product customization and the supply networks complexity have increased and, on the other hand, the product life has shortened.

As shown in [Lee04], where more than 60 leading companies have been observed, a successful supply chain muss not only be fast cost-efficient but it must be also designed following the triple-A concept, i.e. agility, adaptability and alignment. Those last three characteristics were found to be key factors for a durable and well-designed supply chain, which must be able to survive in the modern competitive context. For that reasons, many resource works of the last years have focused on the mitigation the impact of uncertainties and risks by proposing robust policies for the supply chain. As explained by [Li13], a robust system can involve three levels of control activities, namely operative, tactical and strategic. The first one refers to actions which can be implemented in the short term but they generally have small impact on the system, namely that, if the production context change is huge, their effect is not enough to solve the problem. In that case some tactical or, if a radical reconfiguration is needed, strategic activities can be carried out to make the supply chain remain competitive.

Some researches focused only on the strategic ones ([CS04] ; [KS05] ; [Tan06b]), while others on the tactical ([JG95] ; [GT03] ; [TW05] ; [Tom06] ; [HIX10]) or operational activities. Very seldom two or three groups of policies or different control activities are considered together, like in [Li13]. For instance, [SD06] considers only the facility location (strategic) and [CL05] and [MdASL05] consider only the material procurement or revenue sharing contract (tactical).

According to [Tan06a] and [Li13], it is possible to classify the different reaction activities not only according to the implementation time and effort (operational, tactical and strategic) but it also possible to divide them in 4 groups:

- Production management : flexible production rates (e.g. machine service time and number of servers), dynamic inventory control (e.g. Kanban number)

- Supplier management : choice of the main and backup suppliers, urgent deliveries
- Demand management : dynamical pricing, product substitution
- Information management : information sharing among the supply chain partners (e.g. current inventory levels, production status, customer demand rate), collaborative planning

It is important to notice that in each of those groups there can virtually be at least one control policy for each of the three levels and vice versa.

[Tan06a] also stated that supply chain risks can be divided into groups: operational and disruption risks, according to the frequency and the impact on the supply chain. The first ones refer to frequent risks which define the normal system variability such as typical machine failures, small organizational problems, daily demand fluctuation and so on. Those risks have usually a small impact on the system performances. By contrast, typical disrupting risks can be a systematic change in the mean demand, a serious failure in the plant which blocks the production for days, decreasing of the delivery reliability but also nature disasters and economic crises. Those events have a huge impact on the supply chain and some stronger reconfigurations are needed.

3.2 Development of lean production

The risk management is one of the central topic when we are dealing with lean production systems. The first idea of building a lean system was in the Toyota production system (TPS) ([Mon83] ; [Ohn88]). After the positive experience in this company, the topic was further investigated in [Gro93] and [Spi90] with the Just-in-time philosophy derived from the TPS.

Later on, in the western industries, the concept of Just-in-time evolved in the "lean philosophy" with great attention also to the hidden costs, efficiency in the material flow, decreasing waste ([CMMT05] and [RKTT11b]), minimizing the work in progress, flexibility, customer service level, low setup times and so on ([Kra88] ; [WJR90] ; [Zip91] ; [AG07]). In order to achieve all those goals at the same time, the use of the "lean philosophy" is the only possible way ([BTS92]). The advantages that the lean production can potentially bring to the company have been widely studied in plenty of other works ([SB11] , [CCS04]).

Furthermore, [RKTT11a] investigated how lots of western companies showed interest in this new production methodologies and how they adopted Just-In-Time philosophy. Moreover, [PST10] stated that in a typical western company approximately between 40% and 70% of the activities are not strictly necessary and they could be eliminated along with their costs. As a result, the use of the lean techniques implies a much higher competitiveness against the plant relocation in countries where the manpower costs less. The typical set of lean production tools are cellular manufacturing, pull mechanism for the stock management, total quality management ([BRA94]), rapid and frequent setups, production leveling, etc.. The tool, on which the research has focused the most in the

last years for this kind of systems, is for sure the pull mechanism and, in particular, the one implemented by Kanbans. If the pull and push mechanism are compared, it can be concluded that the first one is easier to implement because it does not require forecast information, the inventory level can be better controlled and limited and it is much easier to detect problems. By contrast, it has also some disadvantages. For instance, it requires a cultural change in the company and it does not work well in an unstable environment.

The typical issues concerning the lean production systems are the determination of the best parameter setting for a certain scenario, the performance comparison between pull techniques, applicability of a lean technique in a realistic problem and so on. Those topics have been investigated, for instance, by [Ber92], [DG92], [HK96], [AE99], [KP07], etc.

The modeling approaches can be divided in three main categories:

- Deterministic models (mathematical programming, Toyota formula, etc.). For example, [BC87] formulated a mathematical programming model to determine the optimal parameter configuration for a deterministic multi-stage assembly-structure Kanban system and [BDMF01] treated the Kanban system as a multi-class queueing network and developed approximation methods for more general Kanban systems

- Stochastic models (queueing theory, Markov chains). [Buz89] developed a linked queueing network model to describe the Kanban system behavior. Moreover, other works, such as [DHMMO89] and [AGMG93], modeled the system with Markov chains. In particular, the first one developed a discrete-time model to analyze the operation of the Kanban mechanism, while the second one built a continuous time model to determine the optimal number of Kanbans for multi-stage multi-product case. An other relevant study with stochastic models are [FDMD95], which presented a queueing network with a synchronization mechanisms for a single-product multi-stage serial line Kanban system and proposed an approximation method to determine the optimal parameters at each stage by considering each section as a standalone sub-system.

- Simulation models. When the studied model is too complex and its simplification would bring to wrong or useless results, the simulation approach is the only solution. This class of models is generally used to compare different pull techniques or when there are many control factors ([PB96] , [HRT83] , [KKRW87] , [PRTIH87], [BCG97], etc.). For instance [KKRW87], conducted comprehensive experiments based on simulation models to investigated the impact of various control factors like setup times, lot size, production rate, worker flexibility and production structure in Kanban systems. An other relevant example is [BCG97] which compared different pull control mechanisms like base-stock policies, CONWIP and hybrid Kanban-CONWIP through simulation.

The success of the TPS has triggered the development and the improvement of the

lean techniques. Nowadays, it is possible to find plenty of different control strategies in the literature and in [Li13] they are summarized as follows:

- K : traditional TPS Kanban systems
- K+ : modified Kanban systems (CONWIP, CONWIP-Kanban)
- K+B : extended and generalized Kanban control systems, i.e. policies which combine the Kanban and the base-stock policy.
- K+time : reactive and adaptive Kanban control systems
- K+B+time : extended-CONWIP-Kanban control systems
- K+time+service rate monitor : robust Kanban systems

The K letter refers to the traditional Kanban policy used in the TPS. The B means "Base-stock policy" which is derived from the traditional inventory control models ([CS60] ; [Kim88] ; [Axs07]), namely that the inventory level at the final stage is monitored and it is kept within a specific range as much as possible. In particular, when the inventory level becomes dangerously low, the upstream part of the supply chain is triggered to fill the gap. The K and B policies represent the two main tools for the pull mechanism and each of them has advantages and disadvantages with respect to the other one. The K policy assures that the inventory levels remain limited and it is easy to be implemented. By contrast, the B policy is more robust when the environment is uncertain and the demand has a greater variability because the information can be delivered quicker along all the supply chain and not only between adjacent stages. Moreover, as [LPW04] stated, if the inventory control decisions are made only based on the demand of the immediate downstream stage, the Bullwhip effect is usually experienced. In order to increase the response time [SWH90] proposed a CONWIP (Constant work in progess) model, which consider the whole system as a unique stage with a unique Kanban cycle. It means that, when the cards are detached from the finished products, they return immediately to the first station and not to the immediate upstream station. However, this systems have some shortcomings which are typical of the push techniques and the problem of the information delay still remains.

In order to exploit the advantages of both the K and B policy, some hybrid approaches have been studied denoted with K+B. The first two hybrid models were the extended ([FDMD95] ; [DL00])and the generalized Kanban control system ([Buz89]). In those two models the production at each stage is controlled by two parameter : the Kanban number and the target inventory level. However, they present some differences. Particularly, the control mechanism of the first one is less complex and the demand information is directly transmitted to each stage as a global information. By contrast, the generalized Kanban control system hasn't this feature but there are no constraints on the number of Kanbans at each stage as in the previous one.

If the performances of traditional lean systems with K and B policies are compared with the ones of hybrid systems, it is possible to find out that the latter does not

always perform significantly better. Indeed, if the environment is not turbulent, the traditional ones are simpler to be implemented and managed and perform as well as the more complex ones. On the contrary, if the environment is uncertain and turbulent the hybrid ones can achieve better performance due to their robustness. Those comparisons can be found in many works such as [BCG97] , [DFDM00] , [KD00] , [LD00] , [GH05] and so on.

3.3 Adaptive and reactive lean production systems

Until now, only systems with a configuration which is optimized and kept fixed have been considered. However, it is also possible to design a system which monitors the current system state and can be adapted if considered necessary.

In this field of research an important contribution was given by [CGT08]. This work deals with the design of an intelligent system which monitors its status and is able to be reconfigured or rescaled when required. The system monitoring is carried out mainly with intelligent sensors which can properly detect problems. In [CFT$^+$10] and [MFG$^+$11] sensors are also used to properly detect a risk situation and to respond to it. In particular, they deal with the so-called condition based maintenance which can be applied when the machine failures are predictable by monitoring some signals or indexes and, in this case, the mean value and the variance of the absorbed electric power can detect a malfunction or an impending failure of some components.

In other cases, the monitored parameters are related to the overall system performances. For instance, in [MGG10] it is suggested to monitor a parameter denoted as OSE (operating system effectiveness). This parameter can detect the main causes of low performances and indicate where the system can be improved. In this work it has been demonstrated how this method can improve the system performances.

In general, if a control policy which adjusts the system is well-designed, the performances are always better. This can be also seen in [CKT09] and [PJT13] which propose a control policy used to a system more energy-efficient. In particular, in both works it is proposed a method which monitors the system state, analyzes it and manage it dynamically.

Some other important contributions have been given by [TN99], [Tak03], [TM01] and [Boo05] with the concept of adjustable parameters during the system operation. In the first three researches the number of Kanbans at each stage can be controlled and modified to improve the system performances. The basic idea is that, when the environment is repetitive and not turbulent, the system can operate with a small number of Kanbans to limit the WIP, lead time and, as a consequence, the overall costs. By contrast, if the environment becomes suddenly turbulent, some additional Kanbans are added because a little increase of the stocking costs is justified by the fact that a larger mean inventory levels are needed to maintain a high service level, which is generally a priority for the company. This kind of systems are denoted "K+time" because the parameter "Kanban number" is used to control the system but it is also changed dynamically as reaction policy. In those works it is demonstrated how the reactive policy improves the

performances especially in turbulent environments.

The main difference between the three approaches is the monitored parameters. In [TN99] the demand data series is monitored and an exponentially smoothed demand interval time is computed and used as controlled index. A lower and an upper level are defined through the concept of control limits from the traditional statistical control charts and, as soon as the controlled index exits the normal range, the Kanban number is modified. On the contrary, in [Tak03] a model called inventory-based reactive Kanban system is presented, where the monitored parameter is the smoothed and scaled inventory level of finished goods. Also in this case, an upper and a lower bound, which contains the normal range for the index, are defined and the system reacts as soon as the index exits the normal range. Moreover, in [TM01] a slightly different model, whose name is inventory-based adaptive Kanban control system, is proposed. In this case, the system can adjust the Kanban number only upon each demand arrival depending on the inventory level.

Furthermore, in [Boo05] it is presented an Extended-CONWIP-Kanban (ECK) control system and, in particular, it uses Kanban, base-stock and CONWIP policies simultaneously to control the material flow and both the base-stock level and the number of Kanbans are designed to be adjustable during the system operation. For that reason, the model is not only denoted by K and B but also by "time", as a consequence, it belongs to the group "K+B+time". With such an adjustable mechanism the system becomes more flexible and robust and, as a result, as demonstrated in [Boo05], it performs better (higher service level, lower inventory level, limited impact of uncertainties on the performances, etc.). Nevertheless, this model has two main limitations. Firstly, it can only model demand-side risks but, in practice, the risks are also connected to the suppliers, the production plant, the manpower, etc. Secondly, it considers two parameters which, in case of relevant and unexpected problems, are not useful to solve the problem.

The last and more recent example of robust Kanban system design is presented in [Li13]. In particular, the work considers a single-product serial-line Kanban system with 5 stages, where the first stage represents the supplier and the last represents the customer. In the model there are many different sources of variability, i.e demand fluctuations, machine failures, stochastic production rates, stochastic transportation times and unreliable suppliers at the same time. Some indexes, which are related to the demand rate divided by the production rate of the system bottleneck and to the inventory level of finished products, are defined and their current smoothed values are monitored. When they reach dangerous values, for example when the demand becomes to big with respect to the bottleneck production rate or when the inventory level becomes too low, it is estimated if it is economical advantageous to modify some system characteristics to make the system return in a safe operation condition, i.e. without the risk of big stockout or inventory costs. Each domain of the two indexes is divided by thresholds in different ranges and each of them identifies a precise reaction procedure to be followed in case that index is inside it. This model includes both the supply and the product management reaction policies, which also contain all the three control levels (strategic, tactical and operation). In particular, it is considered that the more dangerous the actual situation is (for example, the more the inventory level is far from the safe range),

the more effective and expensive the reaction that is taken into consideration is. For example, if the inventory level is a little lower than the safe range, an operational reaction will be considered but, if it is dangerously low, a strategic one will be considered instead because, although it costs more, it avoids some bigger costs like the ones related to stockout and backlog, which are more likely in this second case.

This model achieved important results but it has also some disadvantages. First of all, during the design of the reaction policy it is chosen to find a reasonable solution instead of the optimal one. The reason is that, since many escalation thresholds and index smoothing coefficient should be optimized at the same time, the computation time would be too big for a practical used. Secondly, the model inputs require as huge amount of information (reconfiguration costs for each policy, backlog costs, service time distributions at each stage, etc.) which is generally not available. Lastly, the only considered reconfiguration delay is the one of the review period, which corresponds to the time between two reconfiguration evaluations, but, when a parameter change is evaluated as economical, there is no implementation delay. Indeed, in the reality the delay time of adding a Kanban (operational level) can be sometimes negligible but, by contrast, the longer ones of strategical activities must be considered to obtain accurate results.

The lean production systems which include a reaction policy have been attracted the attention of many researchers and companies in the recent years. The reason is that they have the already discussed advantages in comparison with the traditional ones, namely that they can adapt themselves to tackle the system uncertainties. Nevertheless, they have also many disadvantages. In the traditional system we just have to optimize one configuration for a given environment. By contrast, the robust lean production systems require the user to answer those questions:

- What kind of robust approach should be adopted?
- Which parameters of the system should designed as adjustable?
- What indexes should be monitored to proper detect an impending risk?
- When and how it is economical to reconfigure the system?

The best policy for a defined system and for a defined environment is quite hard to find and generally it is only provided a reasonable solution but not the optimal one. Even if we would be satisfied by a reasonable solution, the optimization generally requires a large amount of data and big computation times especially using simulation. For that reason, the convenience of exact models, which require a small amount of information as input and are usually faster, becomes even bigger for this subset of manufacturing problems.

3.4 Concept of thresholds

As already said, the basic idea of systems which can change their parameters to adapt itself to environment variations is that some indexes must be monitored and depending

on their current values the system can adapt itself to respond to impending risks. Each index domain is generally divided in n ranges by $n-1$ thresholds which define the limit between two regions where different preferable configurations can be assigned. As a result, when a threshold is overcome and the current configuration is no more the preferable one, the system will be reconfigured as soon as possible.

In the literature there are many examples where the threshold concept was used. For instance, for what concerns the base-stock policy, the inventory in the output buffer of each stage is controlled within a specific range and, as soon as the current inventory level becomes dangerously low (below a certain threshold), the upstream stages are triggered to produce more. More in general, the behavior of the upstream machine depends on the current level of the buffer. This approach can be used in the modeling of hedging point policies ([Ger00] ; [KG83]), energy saving policies, restart policies ([GGF$^+$09] ; [GGF$^+$13]) or bulk arrivals or batches ([CG10]), in which the upstream machine is not allowed to start a new batch if the buffer level is above a certain threshold which depends on the batch size.

This concept can be also used to model the behavior of loops and multiple loops, where a fixed amount of parts circulate in a closed cycle. That kind of system generally models pallets and fixtures or Kanbans in control systems ([Ohn88]) and CONWIP ([SWH90]).

[Mag00] and [MMGT09] presented a new decomposition method to model closed loops. It is based on the decomposition presented in [TG98]. The proposed method returns accurate results because the correlation between the number of parts in the buffer and the total population in the system is considered. However, it is not usable for more than 3 machines. Moreover, in [GMM$^+$01] , [Wer01] and [GW07] two important phenomena related to closed loop systems are identified, namely the range of blocking and the range of starvation. The use of the traditional decomposition ([TG98]) encounters some problems in the modeling of those two phenomena because it assumes that the blocking and starvation can propagate along the whole line. The author tackled the problem by determining the range of blocking and the range of starvation and by defining a proper threshold for each buffer. Since at that time there was no model capable of modeling a two-machine line with thresholds on the buffer, the threshold was eliminated by splitting the buffers in two parts in correspondence of it and inserting a perfect reliable machine. However, with this methodology some issues are experienced, i.e. a discontinuity of the production rate as a function of the loop invariant, named "Batman effect" ([SG09]), which is disruptive during the optimization phase.

The first model which uses the concept of threshold in the building block was presented in [TR13]. The idea of building block was first introduced by [GFF07], where the authors defined a discrete-time model of a two-machine line system with finite buffer capacity and with whatever number of up and down states for the machines. Later on, a continuous-time model, which treats the produced parts as a continuous material flow, was presented and developed in several works ([TGYG07] , [TG09] , [TG11] , [Tol11]) but solved differently.

In [TR13] the multi-threshold model for the performance evaluation of a building block with general thresholds was presented. The basic assumption of the model are:

- two-stage production line separated by a finite capacity buffer
- the upstream stage is never in starved and the downstream stage is never blocked
- The flow of processed material resembles a continuous fluid, which flows at a rate defined by the machine rates of the current states
- Continuous-time mixed-state Markov chain
- The behavior of the two machines can be described by whatever number of up and down states and it can be different for each buffer range lying between two thresholds and on each threshold.

The exact methodology is explained more in details in chapter 5. However, the model returns the steady-state probabilities and they can be used to compute the system performances. This model seems to be suitable to study and design robust lean production systems which monitor the inventory level of finished products, react by changing some characteristics of the production system (upstream stage) and include an uncertain customer behavior (downstream stage). The machine behavior must be modeled by means of continuous-time Markov chain. Furthermore, it can also model the reaction delays, namely that, when a new configuration, which is different from the current one, becomes preferable, the system doesn't carry out the reconfiguration immediately but with a certain delay distribution.

3.5 Modeling of the production rate distribution

The behavior modeling of manufacturing systems is a key problem when a certain system must be studied and/or optimized or when we want to investigate the impact of some parameters on its performances. Many methods have been developed in last decades to investigate the dynamic behavior, estimate performances and supporting an efficient design, improvement or reconfiguration. Those methods are generally based on simulation, analytical tools and operating curves ([NvCFF05]). Most of them investigate only the first order performance measures, such as the average throughput, average Work in Progress (WIP) and the average lead time ([KHW98]) but, typically, the higher order performances are not measured ([Hon05]) and the analysis of production variability is still in its earlier phase.

Since there are usually some variability sources in the production systems, for example due to machine failures, randomly distributed service times and so on, the higher order performances are relevant to correctly predict risks and take the right decisions. For instance, the company could be interested not only in the expected production volumes in the next time period but also in the probability of not satisfying the customer demand. Indeed, there is an industrial evidence that the production variability may drastically impact and compromise the performances.

As already said, only few works have already focused on the output variability and only for simple cases, such as for a two-machine system ([Tan00]), for unbuffered lines

([FSHK03]) or for systems with reliable machines ([HWL07]). There are only few examples in the literature, where more general methods have been proposed to evaluate the higher order performances. [CMT10] presented a theory and a methodology to analyze the production rate variability in unreliable manufacturing systems modeled with discrete-time Markov chains and, in particular, the dependency of the variance on the system parameters was investigated. Later on, [ACM14] proposed a general methodology to analize the variability of the output of unreliable small-scale multi-stage systems modeled as general markovian structures. The approach is based on the autocorrelation structure of the system production. At last, [AHC14] analysed cumulated output and lot completion time moments of Markovian reward models where both continuous and discrete cases are considered.

The methodologies and formulas presented in those studies can help the design of a Markov chains, which approximate the production rate behavior of a given system. The same problems are investigated in [HT02] which presented the Ph-Fit tool, i.e. a general phase-type fitting method. This tool receives the production volume distribution in a defined time period as an input and returns a Markov chain which approximates it with a phase-type distribution. The approximation can be decreased at will but the number of required states will increase and it implies that the computation time increases as well.

The research areas presented in this chapter are all related to this master thesis and, in particular, it has been decided to use the multi-threshold model to investigate the chosen set of problems because:

- It allows to represent different production behaviors depending on the inventory levels
- The switching delay between two configurations can be modeled
- The required performances for the cost evaluation can be obtained by elaborating the distribution functions and the mass probabilities which are returned as an output
- The fact that it is an analytical tool implies suitability for a decision parameter optimization and a limited computation time

4 Methodology

This chapter focuses on the general methodology to deal with the chosen subset of systems defined in the problem statement section and how the developed model can be used in the decision-making.

The first part defines the reference situation which must be modeled and studied. Secondly, the multi-threshold model presented in [TR13] is proposed to solve the problem and its assumptions, characteristics, required inputs and provided outputs are briefly recalled. Thirdly, the general modeling to study the chosen subset of problems is presented and it is explained in detail how a number of possible upstream machine configurations can be modeled and linked to the buffer ranges. Precisely, each buffer range must be related to a different preferable system configuration in such a way that the upstream machine is reconfigured when the inventory level becomes dangerously low or big. Furthermore, it is also explained how to model a typical demand behavior and the fact that, when the buffer is empty, the demand is not lost but it is always entirely satisfied. Fourthly, the procedure to obtain the system performances and, then, to convert them in costs through a cost function is explained. In the last part, it is said that, once the model is able to evaluate the costs for a given reaction policy (identified by the threshold position), an optimization algorithm can be used to find the optimal reaction thresholds for a given production environment and the optimal reaction policy can be then used for the decision-making.

4.1 Description of the system to be modeled

For the sake of simplicity, it is decided to deal with a single (or equivalent) product and, since the last part of the supply chain must be design to absorb the customer fluctuations and provide more stability for the upstream departments and suppliers (very important in a lean production system), it is decided to focus on the final part of the supply chain, namely that only the customer, the warehouse of final products and the immediate upstream production stage are considered. The production stage can be configured in different ways, which has a different stochastic production behavior. The possible system configurations are identified by a set of adjustable parameters which concern the production plant and the backup supplier(e.g. production planning, number of servers, extra shifts, extra supplier activation, etc.). Moreover, the production stage is unreliable due to many uncertainty sources such as the component shortage (which is generally caused by the supplier unreliability), internal failures, lack of manpower, manpower efficiency variability, quality problems,etc. Furthermore, when the buffer becomes full, the production stage stops producing until some more parts are removed.

The warehouse level is evaluated every assigned review period and used as a monitored index to detect the risk of backlog, high stocking costs and production stage blocking. Some control ranges for the inventory level are defined and a preferable system configuration is assigned to each of them to react to potential risks. As a result, the number of index ranges is equal to the considered configurations. Only at the beginning on each review period a reconfiguration can be carried out, if the current production configuration is not the preferable one according to the inventory level. The reconfiguration actions can be implemented with a certain delay different from zero. For what concerns the customers, they remove products from the final warehouse with a certain behavior which is independent from the inventory level and, if the buffer is empty, they retrieve the required products as soon as they are available and no demand is lost.

The cost-efficiency of the system will be evaluate by considering the following costs:

- the inventory costs
- the backlog costs
- the configuration costs, i.e. every configuration has a different operative cost

Such a situation can be properly modeled and studied with the multi-threshold model presented in [TR13].

4.2 Multi-threshold model

4.2.1 Conceptual model description

The detailed explanation of the model is presented in the paper [TR13]. However, some general characteristics, assumptions and considerations are briefly recalled in this paragraph.

4.2.1.1 Assumption and description

The modeled system consists of a two-stage production line decoupled by a finite capacity buffer. The flow of the processed material resembles a continuous fluid, which flows at a rate at which the machines operate. The system is modeled as a continuous time, mixed state Markov chain whose states can be represented in the following way (x, M^u, M^d). The x variable represents the amount of material in the buffer and M^u and M^d represent the states of the upstream and downstream machines. According to the buffer level, it is possible to distinguish the states of the system into several ranges, that are delimited by thresholds. The states contained within these ranges are the internal ones and they are continuous states. On the other hand, the ones lying on the thresholds are called threshold states and are discrete. As a result, it is possible to model a different behavior for both machines according to the current inventory level.

4.2.1.2 Inputs and output

The model requires as input the following information:

- the T threshold levels (the lowest one is always located in 0 and it doesn't require to be defined), which identify T ranges
- the behavior of both machines in all ranges is defined in terms of continuous-state Markov chains
- the states on the upper and lower threshold, i.e. the zero threshold and the T^{th} threshold
- the additional threshold states, if it is possible that, when a threshold is reached, the system reacts and makes the rates of both machines equal. Only in this case some additional on-threshold states can exist
- the behavior of the system, when a threshold is reached from whatever internal state in terms probability to cross it and reach a internal state on the other side or to remain in a threshold state or to return in the same range as before
- the behavior of the system, when it is on a threshold state, namely the transition rates from each on-threshold state to all the other states belonging to that threshold and to the adjacent ranges

Once all the information is provided, the states on each layer and threshold are divided in 4 sets, i.e. the γ set (if $\mu_u > \mu_d$), the Φ set (if $\mu_u = \mu_d \neq 0$), the Ψ set (if $\mu_u = \mu_d = 0$) and the Δ set (if $\mu_u < \mu_d$), the parameters $U^u_{i,r,t}$, $U^d_{l,r,t}$ and $\gamma^{r,t}$ are computed for each layer and finally the $C_{r,t}$ and the mass probabilities are computed by using the boundary conditions. The probability distributions for the internal states can be now calculated with the formula 4.1.

$$f(x, \alpha^u_{i,t}, \alpha^d_{l,t}) = \sum_{r=1}^{R_t} \cdot C_{r,t} \cdot U^u_{i,r,t} \cdot U^d_{l,r,t} \cdot e^{\gamma_{r,t} \cdot x} \tag{4.1}$$

The model returns also the mass probability for each threshold state. They can be then elaborated to obtain some mean information about the steady state condition such as mean throughput, mean inventory level, lost demand, probability of being in a certain range or on a certain threshold and so on.

4.3 Modeling of the considered system

The modeling of our system does not include any threshold states, except for the ones on the extreme thresholds, because it could happen only if the decision delay was zero, i.e. the reconfiguration is simultaneous with the arrival on the threshold, and it was possible to have the same rate on M^u and M^d at the same time. Although the second condition

is possible, the first condition is generally not true in a real production system because there is always a decision delay different from zero, even if it can be extremely small. However, in the studied subset of problems it is decided to leave out this possibility. As a result, when a threshold is reached, it is always crossed without changing the state because, since the production rates do not vary in that instant and other transitions cannot happen simultaneously, the production rate difference doesn't change as well and, if the buffer level was decreasing, it keeps decreasing and vice versa. It also implies that the set of internal states is always the same in each range although the transition rate topology varies because the system tries to reconfigure differently in each range.

4.3.1 Threshold number and negative inventory level

Since we want to assign one preferable configuration to each index range and there are C possible configurations for the production system, C buffer thresholds, except for the already defined x_0 threshold, are defined. Then, it is needed one additional threshold to model the negative inventory level, namely that there are actually $C + 2$ thresholds (from x_0 to x_{C+1}) but the buffer level which corresponds to neither stored nor backlogged parts is x_1 and not the x_0. As a result, when the x_1 threshold is reached from above (for a practical point of view it means that the buffer has been emptied), the M^d keeps removing products from the buffer and the buffer level becomes negative. When the production rate becomes again greater than the demand rate, the backlogged demand can be recovered without being lost. In order to model the fact that the demand is never lost, the width of the lowest range, i.e. the one between x_0 and x_1, must be theoretically infinite but, since it can not be represented by the model, a sufficiently wide range is used instead and it is checked that the probability of being on the x_0 threshold is negligible. Furthermore, it is hypothesized that the preferable configuration on the upper threshold is the same as the one of the upper layer $N^{th}layer$ (see 4.1) and the preferable one of the lower threshold x_0 and of the negative layer, i.e. the first layer, which in the reality correspond both to an empty buffer but with different backlogged quantity, is the same as the one of the second layer.

A general representation of the layers and the thresholds is provided in figure 4.1.

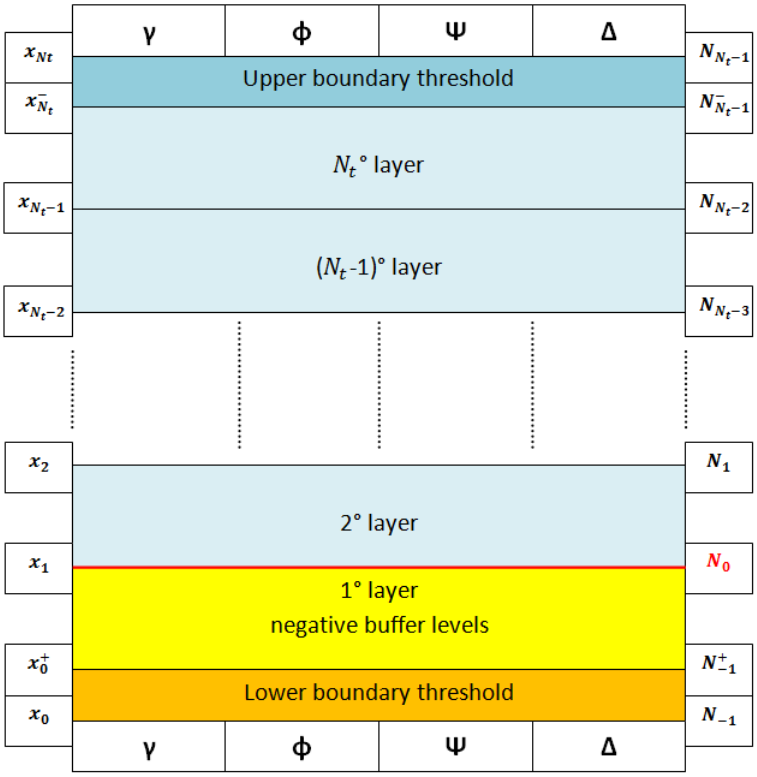

Figure 4.1: Layers and threshold representation

The threshold corresponding to the zero buffer level is the red one and the negative layer and the lower threshold are represented in yellow and orange respectively.

4.3.2 Upstream machine - production system modeling

For each index range, including the negative one, a layer which represents the behavior of both machine can defined. Since the behaviors of the two machines are independent from each other, the two Markov chains can be also defined separately and a simple algorithm to combine them can be developed. In order to make the explanation as clear as possible, we will present the two Markov chains separately.

For what concerns the upstream machine, the modeling of all the C possible configurations must be included in each layer because they do not depend on the inventory level. The only difference is that in each layer a different configuration is preferable and it is modeled by the decision delay transitions. The Markov chain layout in a situation of 4 possible production configuration can be as follows:

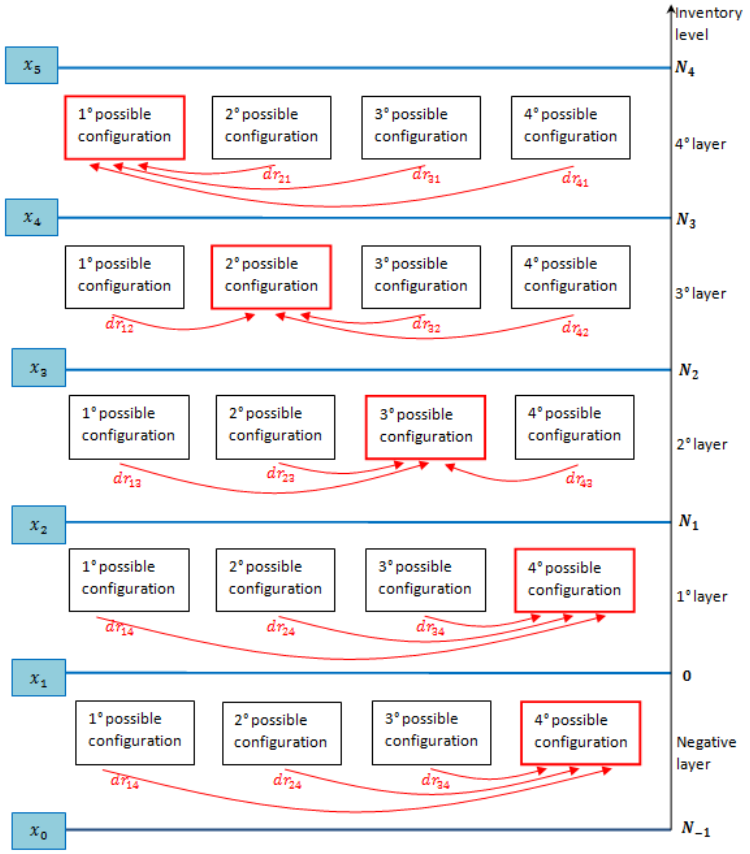

Figure 4.2: Delay transitions and configurations for each layer of M^u

Each square contains a sub-Markov chain, which represents the behavior of the corresponding configuration. It is basically modeled that, if the system is in a certain range but the configuration is not the preferable one, the system tries to reconfigure with a transition delay rate $dr_{i,j}$ which depends on the current i^{th} and the desirable j^{th} configuration (identified by the red square). As a result, the modeled delay time is exponentially distributed. The Markov chain layouts on the extreme thresholds is not represented but, as already explained they are identical to the ones in the adjacent layers with the only exception that the failure rates in the upper threshold are affected by the slowdown phenomenon because they are assumed as operational dependent failures [LMT03].

The delay transitions, which model the reaction delay, must start from each of the states belonging to the i^{th} sub-Markov chain and the arrivals in the states of the j^{th} sub-Markov chain must be randomized considering their steady state sub-probabilities in the standalone sub-Markov chain. For instance, let us suppose that the i^{th} and the j^{th} sub-Markov chain have respectively 2 and 3 states and the $dr_{i,j}$ is equal to 10 and the sub-probability of the j^{th} sub-Markov chain are $Prob = (0.2, 0.5, 0.3)$. Those probabilities can be easily computed by knowing only the sub-Markov chain which models the j^{th} configuration standalone. In this case, the transition rates are computed as follows:

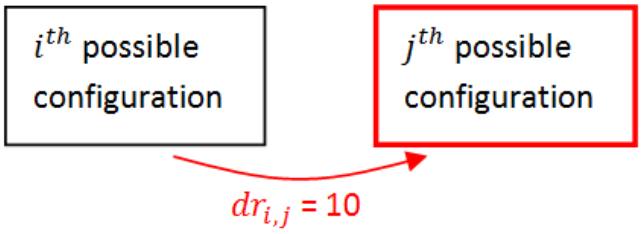

Figure 4.3: Example of delay transitions

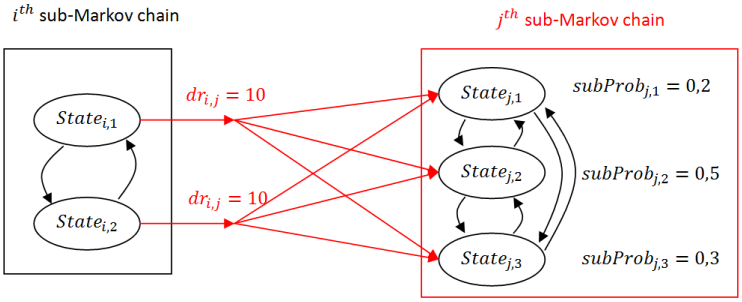

Figure 4.4: Example of delay transitions

The $dr_{i,j}$ must be then shared proportionally to the sub-probabilities of the arrival states :

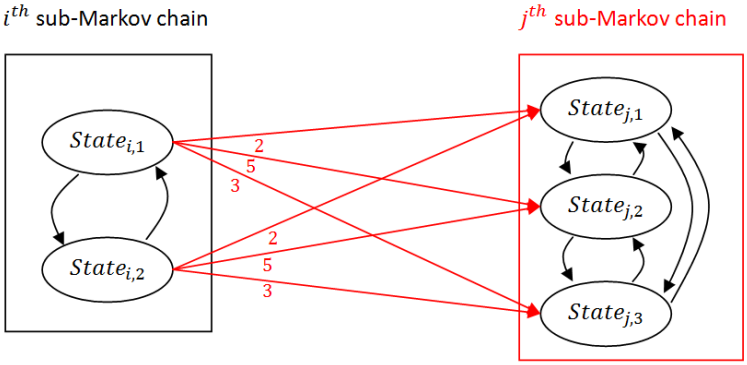

Figure 4.5: Example of delay transitions

On the other hand, if the states of all the sub-Markov chains have a concrete meaning, for instance each state models a precise operational condition or failure mode, and they

are the same (even if the transition rates and the production rates are different), a one to one correspondence can be created. It is reasonable because, for example, if a machine is down in its second failure mode and the system is suddenly reconfigured, the machine remains in this failure mode.

For instance, if both the i^{th} and the j^{th} sub-Markov chain have 3 states, which represent for example the normal production rate and two failure modes, and the $dr_{i,j}$ is equal to 10, the delay transitions are set as follows:

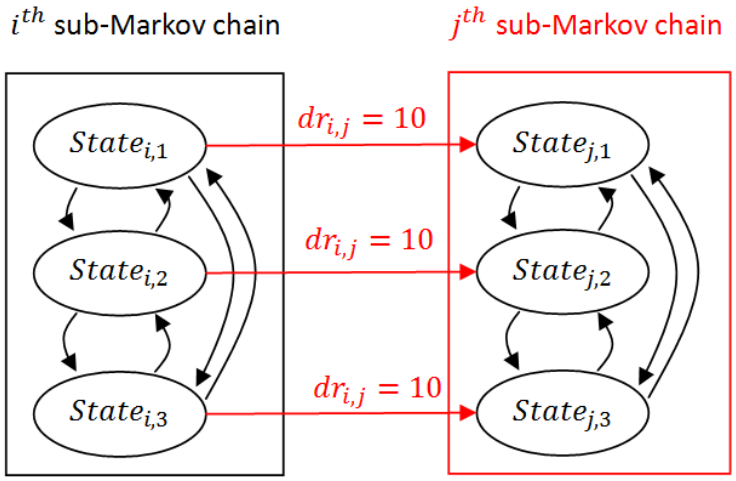

Figure 4.6: Delay transitions with a one to one correspondence

4.3.3 Downstream machine - demand modeling

What is now missing is the modeling of the downstream machine which removes parts from the buffer. The first consideration is that it is reasonable to model a demand independent from the inventory level because the costumer retrieves parts when he needs them and not when the buffer level is high or low and no dynamical pricing policy is considered. For that reason, the following considerations hold for all the layers of the multi-threshold model, namely that the proposed Markov-chain is equal for all of them.

It is assumed that the demand can have different behaviors called scenarios and the transitions between them are described by transition rates $tr_{i,j}$. As a result, the transition times considered in this work are exponentially distributed but it is also possible to use phase-type distribution to approximate more complex ones [HT02].

4.3.3.1 Demand time series

Sometimes the information about the scenario transitions rates or the mean time to transition is not directly available but only the mean probability of each scenario $P_{scenario}$ and the mean time to scenario transition $T_{switch,mean}$ is provided. In this case, it is

possible to model that every exponentially distributed time to transition (with parameter $\lambda = \frac{1}{T_{switch,mean}}$) the next scenario is randomly chosen considering the probabilities $P_{scenario}$. As a result, there exist transitions which link the generic i^{th} to the j^{th} scenario with rate $tr_{i,j} = \frac{P_{scenario,j}}{T_{switch,mean}}$.

For instance, if there are 3 possible scenarios with probabilities $(0.2, 0.5, 0.3)$ and the $T_{switch,mean}$=0.1, the Markov chain for the downstream demand becomes as follows:

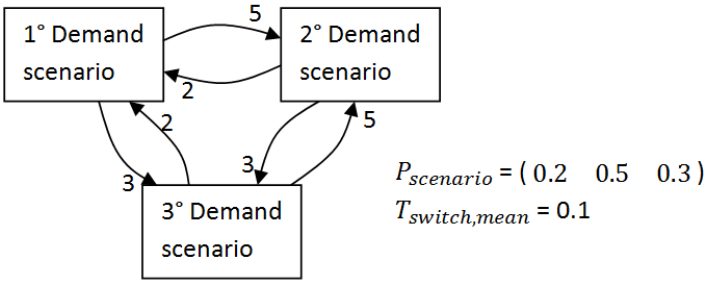

Figure 4.7: Markov chain for independent scenarios

4.3.3.2 Scenarios with multiple states

It is generally useful to represent also some variability included in the scenarios and, as a consequence, more than one state is needed. As explained for the transitions between different multi-state configurations of the upper machine, the same considerations can be done for the multi-state scenario. In particular, if the states of each sub-Markov chain do not represent a physical state which the machine can have but, for instance, they are used to approximate a statistical distribution, the delay transitions must start from each of the states inside the i^{th} sub-Markov chain and the arrival in the states of the j^{th} sub-Markov chain must be randomized considering the steady state sub-probabilities of each of them in the standalone sub-Markov chain.

On the other hand, if the states of all the sub-Markov chains have a concrete meaning, for instance each state models a precise operational condition or failure mode, and they are the same (even if the transition rates and the production rates are different), a one to one correspondence can be created.

4.3.4 upper and lower threshold transitions

For what concerns the states of the upper threshold, there are all states which are also included in the layers except for the Δ states (because as soon as the system enters the Δ set, the system exits from the upper threshold) and the Ψ states (they could be theoretically reached from the γ states when the M_d is already failed, but if it is failed and the buffer is already full, the M_u can't also fail because the failures are operational dependent ODF). On the other hand, all the γ states exist and the threshold can be

reached only through them and the Φ states can also exist but only if μ_u and μ_d can assume the same value and they can be reached only from the γ state sets on the upper threshold. Since the buffer is full when the system state is on this threshold and since the failures are operation dependent, all the rates of transitions which imply the failure of the upstream machine are rescaled by $\frac{\mu_d}{\mu_u}$.

A general representation of the possible states and transitions on the upper threshold is provided in 4.8

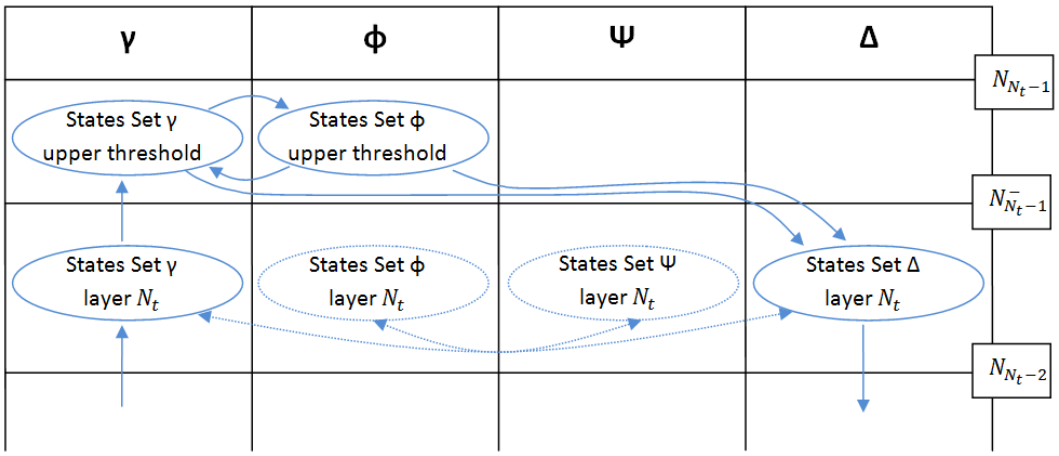

Figure 4.8: upper threshold transitions

For what concerns the lower threshold N_{-1}, it includes the state sets Φ, Ψ and Δ. The threshold is always entered through the Δ states and then, since the behavior of the demand machine is not operation dependent (the demand behaves independently with respect to the buffer level and without the slowdown effect), if M_u was already failed, the Ψ set can be reached. The Φ set can be reached as well if μ_u and μ_d can assume the same value. From each of those 3 state sets the γ set of the first layer is entered as soon as μ_u becomes greater than μ_d and, as a result, the γ state set does not exit on the lower threshold.

A general representation of the possible state sets and transitions on the lower threshold is provided in 4.9

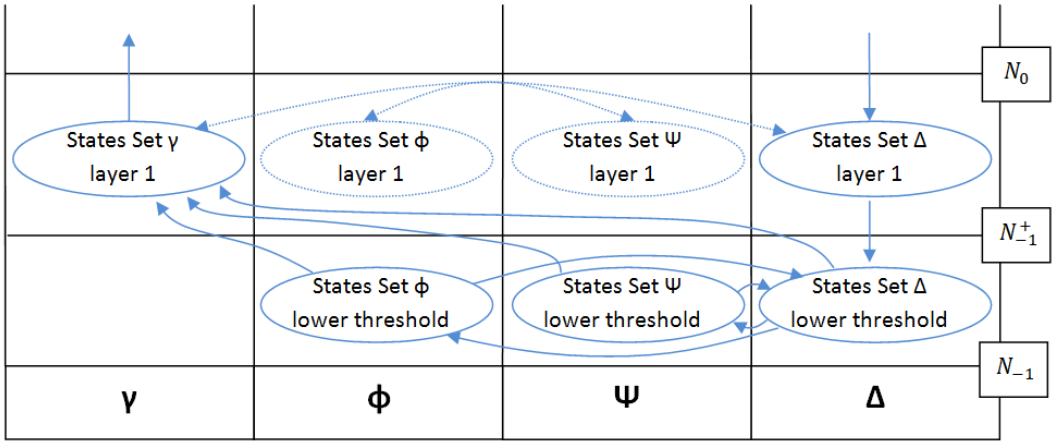

Figure 4.9: lower threshold transitions

Since it is modeled that the demand is never lost, the negative layer must be large enough to make the steady state probability of lying on the lower threshold be negligible. As a consequence the modeling of the system behavior on this threshold has almost no effect on the system performances.

4.3.5 Tools for behavior modeling with Markov chains

If the system has two or more physical states, whose production rate and transition rates are known, there are no great problems in building a proper Markov-chains. On the contrary, if the only information about the system is the probability distribution in a certain time period, some useful tools can help to build the corresponding Markov chain.

4.3.5.1 PH Fit tool

The PH-Fit tool is methodology, presented in [HT02], which is able to find a Markov-chain which approximates a given distribution function. First of all, the considered time step must be defined and then the distribution function, which is given as an input and can be either continuous or discrete. The maximal number of states has to be defined and, in particular, the more states are used, the better approximation can be achieved.

4.3.5.2 Modeling with only 1 up and 1 down state

When the aim is to model a Markov chain with many scenarios on both the upstream and downstream machine, the number of combined states and thresholds along with the computational time can become quite big. If the computational time is a crucial factor for the model usefulness but it wanted to model a production variability, it can be check if the production distributions can be modeled with only one up and one down state. In

some cases, it can be a good compromise but the modeled distribution must be always compared with the real one to assess the approximation.

Indeed, this procedure is not able to control the distribution but it can exactly model the mean value and the variance in a precise time period. The only consideration about the distribution is that, if the rates are big and, as a result, the system changes its state many times in the time period, the distribution tends to approximate a gaussian one. Furthermore, if the system physically has the two represented states and the time to failure TTF and the time to repair TTR are exponentially distributed, the modeled distribution is exact.

The computation of the 3 parameters which defines the Markov chain, i.e. μ, p and r, is carried out with the following method. First of all, it is assumed that the nominal production μ rate is known or, in any case, it is set to a reasonable value which represents the highest possible production rate. Secondly, it is assumed to know the expected value and the variance of the distribution for the considered period and those assumptions represent the two constraints we need to determine the two remaining unknowns.

In order to do it, the method described in [CMT10] can be used. It provides a useful approach to compute the exact steady state expected value and variance of the produced parts of multi-stage manufacturing systems modeled with discrete time Markov chains with whatever layout. In our case the Markov chain is very simple and some easy formulas are already provided for such a system. The only problem is that the formulas are valid for discrete time Markov chain, while the sub-Markov chains are based on the continuous time.

Nevertheless, the approach presented in the paper can be also used to compute the exact parameters for a continuous time Markov chain using the following method, which is based on the fact that the smaller the transition probability is, the smaller the difference between discrete and continuous time modeling will be. For example, let us consider a geometrically distributed random variable X_G with probability p in the considered time unit TU and an exponentially distributed random variable X_E with transition rate t and let us assume that p and t have the same value. From the statistics, it is known that:
$E_G = \frac{1}{p}$
$E_E = \frac{1}{t}$
$VAR_G = \frac{1-p}{p^2}$
$VAR_E = \frac{1}{t^2}$

It can be clearly seen that, if p and t have the same value, the expected value is always the same and the variance of X_E is always greater than the variance of X_G.

If a second geometrically distributed random variable X'_G is considered with the same expected value of X_G but a time discretization which tends to zero, the value of p' tends to zero and, as a result, the variances also coincide.

The method is based on three steps:

First of all, the method considers the continuous Markov chain with the real values (first box):

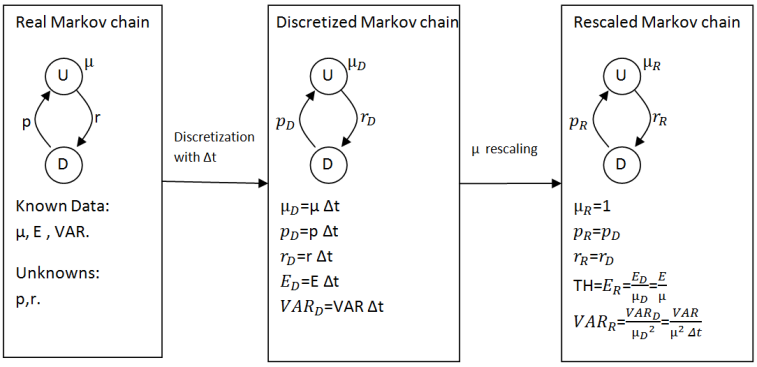

Figure 4.10: sub-Markov Chain

The second step consists in discretizing the time period and considering a very small Δt (time step) to make the transition rates as small as possible. During this step all the rates, the expected value and the variance of the variable are multiplied by Δt and, as a result the configuration of the second box is obtained.

Thirdly, the production rate must be rescaled by its reciprocal to obtain 1 because one of basic assumption of the discrete-time model is that when the system state is up, it produces one part every time step. During this operation the rates do not change but the expected value is also rescaled by $\frac{1}{\mu_D}$ and the variance is rescaled by $\frac{1}{\mu_D^2}$.

Once a configuration where the rates are very small and the up state production rate is equal to 1, the same formulas can be applied to the continuous-time Markov chain by pretending that it is a discrete-time one and the smaller the transition rates, the smaller the approximation. p_R and r_R can be found and then divided by Δt to obtain the real rates.

The formulas provided by the paper [CMT10] are the following ones:

$$\rho = 1 - p - r \tag{4.2}$$

$$TH = \frac{r}{r + p} \tag{4.3}$$

$$VAR = TH(1 - TH)\frac{1 + \rho}{1 - \rho} \tag{4.4}$$

Combining those equations it is possible to define the procedure with formulas and implementing an algorithm which decreases the Δt step by step, providing always a better approximation, until the convergence is reached:

1. A Δt_{start} is defined and it must be very small from the beginning because the corresponding p_R and r_R must be always less than 1 because the will be treated as probabilities.

2. Then the equivalent parameters for the rescaled Markov chain are computed and in particular:
   $p_R = p \cdot \Delta t$
   $r_R = r \cdot \Delta t$
   $E_R = \frac{E}{\mu}$
   $VAR_R = \frac{VAR}{\mu^2 \cdot \Delta t}$

3. Now the formulas presented in the paper can be applied as follows:
   $\rho_R = \frac{VAR_R - E_R \cdot (1 - E_R)}{VAR_R + E_R \cdot (1 - E_R)}$
   $r_R = E_R \cdot (1 - \rho)$
   $p_R = 1 - r_R - \rho_R$

4. After that, it is possible the compute the p and r of the original Markov chain. In particular, $p = \frac{p_R}{\Delta}$ and $r = \frac{r_R}{\Delta}$

5. Finally, if it is the first rate computation, the algorithm goes back to the step 2, otherwise the difference between the previous computation and the new one is computed and if it is lower than a defined value the algorithm returns the actual values, otherwise it goes back to step 2

It is important to notice that equations are not linear and it implies that the solution for a given μ, expected value and variance may not exist. For instance, if the μ and the E are fixed and the variance is gradually increased, the corresponding p and r decrease as a result but at a certain point a solution will no more exist.

4.4 Model outputs

As already anticipated, once all the information is provided, the model returns the probability distributions for the internal states and the mass probability for each threshold state. Then it must be decided which performances and mean indexes are relevant, for example, for the cost evaluation.

The computable parameters of interest of the multi-threshold model are usually:

1. Probability of each internal state $P_{t,i,l}$ (the probabilities of the threshold states are already available as an output without any further computation). In order to compute this probability, the continuous distribution must be integrated. In particular, if an internal state lies on the t^{th} layer (i.e. between the thresholds x_{t-1} and x_t) and it is composed by the i^{th} state of the upstream machine and l^{th} state of the downstream machine, its continuous distribution probability is represented by $f(x, \alpha^u_{i,t}, \alpha^d_{l,t})$ and it must be integrated as in 4.5

$$P_{t,i,l} = \int_{x_{t-1}}^{x_t} f(x, \alpha^u_{i,t}, \alpha^d_{l,t}) dx \tag{4.5}$$

2. The throughput TH, i.e. the mean number of parts which pass through both machines in the chosen time period ($\frac{parts}{timeUnit}$). In our case, it represents the mean number of sold and produced parts at the same time because in steady state conditions they must be equal. Since the downstream machine is never affected by the slowdown phenomenon, the throughput can be obtained by summing all the products of the state probability and the related downstream production rate for all the internal states, upper threshold states (defined by the set Ω_{ut}) and lower threshold states (defined by the set Ω_{lt}).

$$TH = \sum_{t=1}^{N_t} \sum_{i=1}^{I_t} \sum_{l=1}^{L_t} P(\alpha_{i,t}^u, \alpha_{l,t}^d)\mu_l + \sum_{i,j \in \Omega_{lt}} P_{lt,i,j}\mu_l + \sum_{i,j \in \Omega_{ut}} P_{ut,i,j}\mu_l \tag{4.6}$$

3. The mean inventory level $avInv$, which represents the steady-state average quantity of material in the buffer and, as a consequence, it is expressed in *parts*. It can be computed by summing all the state probabilities weighted by their real related buffer level. It means that, for instance, in the negative layer the probabilities are multiplied by zero and not for a negative value which is related in the model. As a result, in the formula the states of the lower layer and of the lower threshold are not considered because they don't give any contribution:

$$avInv = \sum_{t=2}^{N_t} \sum_{i=1}^{I_t} \sum_{l=1}^{L_t} \int_{xt-1}^{x_t} f(x, \alpha_{i,t}^u, \alpha_{l,t}^d)(x - x_1)dx + \sum_{i,j \in \Omega_{ut}} P_{ut,i,j}(x_{ut} - x_1) \tag{4.7}$$

4. The mean backlog quantity $Q_{backlog}$, which represents the steady-state average parts in backlog in a steady state condition and, as a consequence, its unit of measurement is *parts*. This index can be computed in a similar way with respect to the average inventory level but in this case only the states related to a negative buffer level are considered in the formula. In particular, all the probability of the lower layer and of the lower threshold is weighted by its negative buffer level and integrated or summed respectively. As a consequence, the formula will be:

$$Q_{backlog} = \sum_{i=1}^{I_1} \sum_{l=1}^{L_1} \int_{x_0}^{x_1} f(x, \alpha_{i,1}^u, \alpha_{l,1}^d)(x_1 - x)dx + \sum_{i,j \in \Omega_{lt}} P_{lt,i,j}x_1 \tag{4.8}$$

4.5 Objective function

The goal of the model is, first of all, to provide a tool for the differential costs evaluation to compare different settings of the escalation levels and, secondly, to include the model in an optimization tool which is able to find the optimal escalation levels in a given production environment. As a consequence, we must define the outputs which will be considered in the cost function.

In the defined model the throughput will not be considered because it is not related to a differential revenue since the demand is supposed to be always satisfied. However,

the model remains coherent with this assumption only if the probability of being on the lower threshold is negligible, otherwise the demand is partially lost and the revenues would be differential. This condition must always be checked.

Furthermore, we are not interested in the probability of each single state of a layer or on a threshold but, since the operative costs are defined for each configuration, what is important is the probability of being in a certain configuration P_c considering all the states of its sub-Markov chain and all the layers along with the lower and upper thresholds at the same time. Finally, the mean backlog quantity and the average inventory level are considered.

The following formulas it is considered that the costs are expressed in euro and the reference time period is the year. In general they can be different but the formulas do not change. Precisely, the objective function OF, which expressed in $\frac{euro}{year}$, has the following form:

$$OF = C_{inventory} + C_{backlog} + C_{configurations} \tag{4.9}$$

The three cost items in the right side of the equation are also expressed in $\frac{euro}{year}$. By decomposing the three elements of the objective function, it is obtained:

$$OF = avInv \cdot C_{finishedProduct} \cdot r_{stock} + Q_{backlog} \cdot C_{backlog} + \sum_{c=1}^{N_{configurations}} P_c \cdot C_c \tag{4.10}$$

The first product considers the fact that if the products are produced in advance, they must be stocked and it implies some real costs and some opportunity costs. The first ones concern the costs related to the obsolescence and insurance of products, the management costs of the warehouse and so on. The second ones concern to the fact that, if a certain amount of capital can't be reinvested because it consists on stocked parts, the company renounces to the potential corresponding financial returns. Particularly, $C_{finishedProduct}$ is the industrial cost associated with a finished product which is stocked in the final buffer and it is expressed in $\frac{euro}{part}$. If it is multiplied by the average inventory level expressed in $parts$, the average unavailable capital related to the stock (expressed in $euro$) is obtained. r_{stock} represents the yearly cost which the company incurs, if an amount of goods of unitary value is stocked in the final warehouse and it expressed in $\frac{1}{year}$. When those three quantities are multiplied, the inventory costs ($\frac{euro}{year}$) are obtained.

The second element represents the backlog costs. As a matter of fact, if a customer wants to take some parts from the warehouse and he can not because the required product is not available and his demand will be satisfied as soon as the products are produced again, the company usually incurs some penalty costs which can be defined in the contract with the costumer or some costs which correspond to the damages to its public image and so on. In any case, in this thesis they will be quantified in term of cost per product in backlog per considered time period and expressed in $\frac{euro}{part \cdot year}$. By multiplying this parameter by the mean backlog quantity expressed in $part$, it is obtained the backlog costs in the time period $C_{backlog}$. It is important to notice that

this formula model a simplified situation because, for instance, it can happen that the backlog costs are not proportional to the backlogged quantities, namely that, if the number of backlogged parts becomes very big, the company can incur some additional fines. Another example is when the backlogged parts are few and no problems are caused in the production plant of the customer for the lack of components. It this case, it is possible that the company incurs no backlog costs. Other examples where this formula approximates the real costs will be presented for the real case.

Thirdly, the configuration costs are considered. In particular, the cost of each configuration in the chosen time period C_c (expressed in $\frac{euro}{year}$) is multiplied by the probability of being in that configuration and by summing all those contributions the overall configuration costs are computed.

The three elements are then summed and the costs associated with the chosen escalation levels are found.

4.5.1 Optimization algorithm

After a methodology to evaluate the costs of a configuration is defined, the minimum of the objective function in a defined production environment can be found using an optimization solver.

In order to choose the most suitable one, the characteristic of the problem must be defined. The objective function is non-linear and the following constraints are applied:

$$x_1 = sizeNegBuffer \tag{4.11}$$

$$x_i > x_{i-1} \qquad \forall i = 2, .., N_t \tag{4.12}$$

$$x_{Nt} \leq invMax \tag{4.13}$$

The first constraint imposes the width of the negative layer $sizeNegBuffer$, the second one that the range sorting is respected and that the width of all the ranges can not be equal to zero, otherwise a layer which does not exist could be modeled. When the optimal escalation levels are obtained, if they include a layer which has the width which is approximately zero, it is suggested that the related configuration will be never used. The third constraint imposes that there exists a limit on the size of the final warehouse, if, for instance, the space dedicated to the stocking of final products is fixed.

As a result, the problem is a non-linear minimization, namely that the objective function is not linear, with linear constraints. For that kind of problems Matlab, which is the used computation software, provides the fmincon solver. In case no gradient is provided, it has 3 algorithm options:

1. Interior-point which handles large, sparse problems, as well as small dense problems. The algorithm satisfies bounds at all iterations, and can recover from NaN or Inf results and it is a large-scale algorithm. An optimization algorithm is large scale when it uses linear algebra that does not need to store, nor operate on, full matrices. This may be done internally by storing sparse matrices, and by using

sparse linear algebra for computations whenever possible. Furthermore, the internal algorithms either preserve sparsity, such as a sparse Cholesky decomposition, or do not generate matrices, such as a conjugate gradient method. In contrast, medium-scale methods internally create full matrices and use dense linear algebra. If a problem is sufficiently large, full matrices take up a significant amount of memory, and the dense linear algebra may require a long time to execute.

2. SQP satisfies bounds at all iterations. The algorithm can recover from NaN or Inf results. It is not a large-scale algorithm

3. Active-Set can take large steps, which adds speed. The algorithm is effective on some problems with non-smooth constraints and it is not a large-scale algorithm

In order to choose the best algorithm it is suggested to use the interior-point algorithm firstly, then run the optimization again with SQP and then with active-set to try to achieve more speed on small- and medium-sized problems. In general it can be said that the first algorithm is the one which works in the wider range of constrained non-linear optimization problems and, as a result it is used as default by Matlab. On the other hand, for some subsets of problems the other two algorithms work better.

4.6 Use of the model in the decision-making

Once the optimal reaction policy is defined, it is known which decision must be implemented every time it is possible to do it. In general, two cases can be distinguished.

Firstly, when the modeled delay represents the fact that as soon as a threshold is crossed, the needed to change the configuration is detected but the implementation of the reconfiguration takes some time. In this case, it is always possible to take reconfiguration decisions but they are actually taken only when a threshold is crossed and the decision to be implemented is identified by the index range after the crossing. Furthermore, the delay distribution can have different expected values because different reconfigurations can be more time-consuming than others

Secondly, the second case is when the reconfiguration is performed as soon as it is realized that the current index range has changed, i.e. with a detecting delay. In this case it is reasonable to assume that all the delay distributions are equal because they usually depend on the monitored indexes, which in the modeled system is only the inventory level.

For example, if the monitored index is only evaluated every deterministic decision time step DTS, a decision can only be taken in certain time instants and the decision will be immediately implemented. As a result, all the delay transition distributions are equal and uniformly distributed with minimum value equal to 0, in case the decision is taken immediately after the threshold crossing, and maximum value equal to the DTS, when the crossing occurs immediately after the decision instant. As a result its expected value is $\frac{DTS}{2}$.

Of course, it is possible to have both situations combined together, i.e. the monitored index is evaluated every decision time step but the reconfiguration takes some time to be implemented. In this situation, the delay expected value can be different, as in the first case, because it is obtained by summing the expected value of the detecting and of the implementation delay. Moreover, the decisions are taken by comparing the current index value at the moment of the decision with the optimal escalation ranges, as in the second case.

5 Bosch real case

In this chapter the Bosch case is introduced and a general information about the production plant and its performances, the products, the implemented operative policies and the customer behavior is provided. By contrast, the company specifically requested not to share some more detailed pieces of information about the precise production site, the exact product characteristics and bill of materials (BOM) and the used production technology.

The chapter is structured as follows:

1. Products description
2. Plant layout presentation and assembly activities
3. Customer behavior and backlog policies
4. Usage of Kanban card
5. Production policies (available time computation, reaction policies and production leveling)
6. Anticipated replanning

5.1 Introduction

The production is located in Italy and belongs to the Bosch GmbH, a leader company in the production of components for the automotive industry and of many other products. The company has included in his production context the lean manufacturing philosophy along with many of its management policies for many years with a great attention on the cost-efficiency of its production. The result is the achievement of a own way to produce, called Bosch production system BPS.

5.2 Products

In particular we will focus on the final assembly department of the mentioned plant. Two types of pumps denoted as CP1 (figure 5.1) and CP1H (figure 5.2) , which elaborate diesel fuel for a car engine, are assembled in that department.

Figure 5.1: CP1 pump

Figure 5.2: CP1H pump

The two products are quite similar regarding the characteristics of the features to be done and the cycle times at each station but the second one (CP1H) is a little bigger and more performing than the other. Each of those two types is available and produced in almost 60 models, since the product is customizable according to the needs of the customer but, within all those possible feature differences, the cycle times needed to produce each part at each station are quite similar (always within the range $\pm 15\%$). For that reason the production data, that are collected to measure the performances and to compute the OEE and other production parameters, are not disaggregated for each single product but all the products are considered to be the same workpiece. Another reason, which justifies that procedure, is that every day the production managers try to do a reasonably big number of setups.

Moreover, within that variety of final products we can divide them in two big categories, which are called Renner and Exoten products. The first ones are the most regularly and frequently required products. It means that they constitute the 70-80%

of the total production quantity and each product in that group is required every week or even every day. For those reasons they are managed with a tradition pull-technique, namely that, when the costumer withdraws a certain number of products from the final warehouse, new Kanbans are freed and they come back at the beginning of the department to let more parts be produced. The products of the second group are called Exoten and they represent the products which are only seldom required. They constitute the 20-30% of the total production and, as a consequence, it would be pointless to manage them with a pull-technique, because it would make increase the inventory costs and the risk that, once they are stored in the inventory, no customer will require them anymore, and they will become obsolete, although the company has already spent money to produce them or, in any case, they stay longer in the supermarket until a costumer takes them. The best solution is to produce them only when required. The products division in those two groups is performed with a traditional ABC analysis which sorts all the products from the one with the biggest production volume to the one with the lowest one. Then the cumulative quantities are computed and the Renner defined as all products, which have a correspondent cumulative value of approximately 70 or 80 % the final cumulative value.

That department takes as an input all the needed components (approximately 40 components, i.e. the shaft, the head, the ring, the casing and other components of the pump) both from other departments of the same plant and from external suppliers. Since each of those components is used in very limited number of variants, because, even if there are 120 different types of pumps, there exists for example only 4 or 5 types of different heads, it is suitable to use a pull-technique to manage them. It is important to notice that the components standardization is a fundamental characteristic of a typical lean production plan but, since the specific needs of each customer are generally very different, the variety of final products is usually quite big also in a lean production system.

5.3 Plant layout description

The considered department can be divided in six major areas. There is the warehouse of all the components, the first and the second assembly line, the quality control area, the painting area and the final product warehouse. The figure 5.3 represents the plan layout.

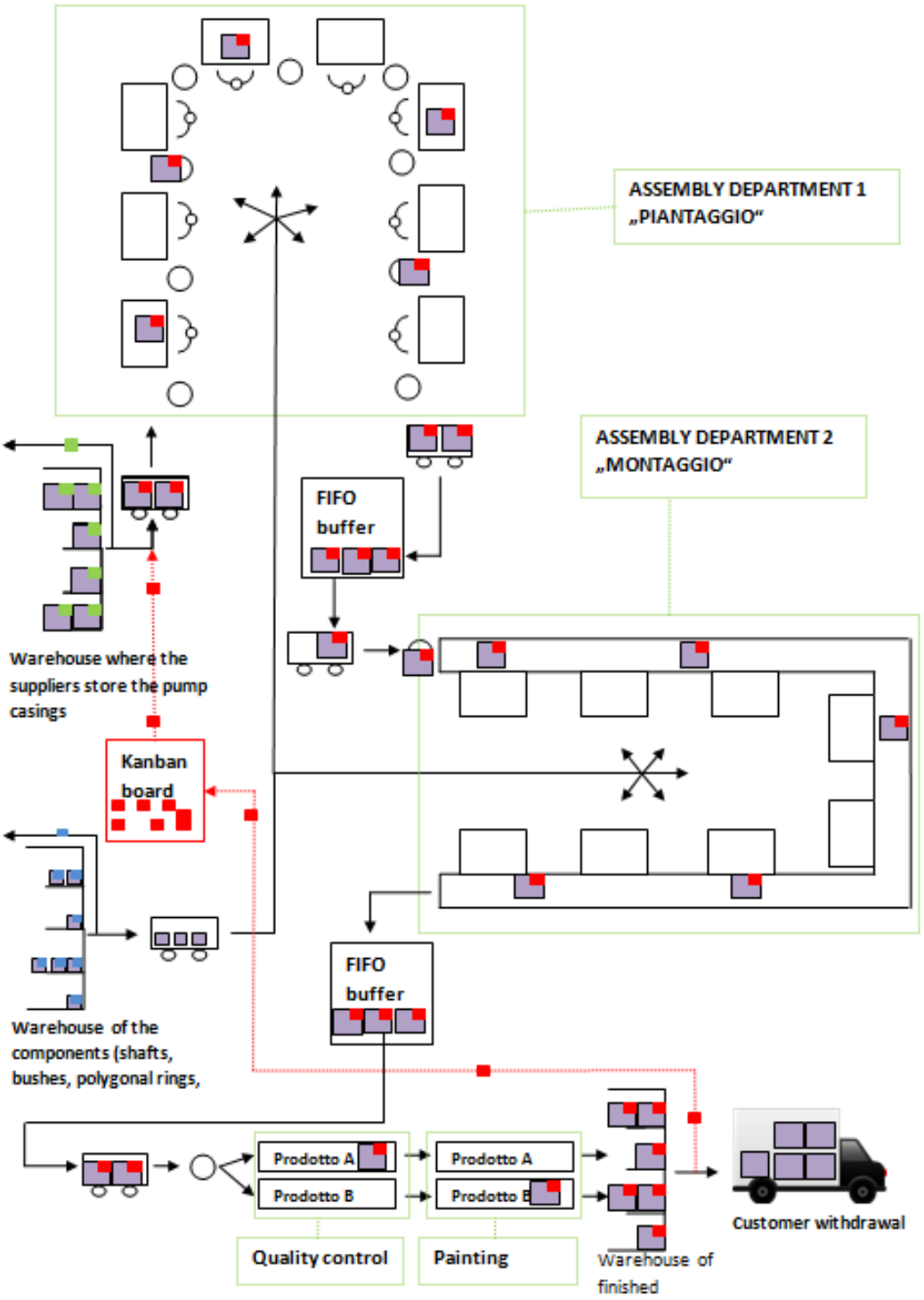

Figure 5.3: Plant Layout

At the beginning of the production line there are the two warehouses. One with all the different types of casing, which are brought there by the supplier and one with

all the different components. Since their number is not so big, and, as a result, each of them is constantly required, all the components are managed with Kanbans (figure 5.4), namely that those cards are attached to all the components in the warehouses and when someone removes them from it, the corresponding Kanbans are detached from the retrieved components and are sent back to the beginning of the upstream line, which produces them. That allows the previous departments to produce new parts. The component is then taken to the station where it will be assembled, either in the first assembly department or in the second one.

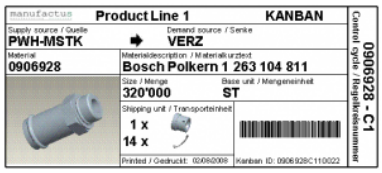

Figure 5.4: Kanban

The casings are considered as the main component and for this reason they are moved through the entire production line. Each station is supplied with the components, that it needs, so that the planned operation can be done and then the assembled casing moved to the next station. Moreover, when it is taken from the upstream warehouse, a production Kanban is removed from the part and sent back to the beginning of its production line. It is important to understand that the assembly department is allowed to take a new casing from the warehouse only when there are free production cards (the red ones in the picture) in the Kanban board 5.5.

Figure 5.5: Kanban board

The first assembly department is composed by 3 separated production lines and each of them is composed of 7 stations which can be coupled together because the buffer between them is small. The operations which are performed are the following ones : the pre-assembly of some basics components, hammering of bushes and of the polygonal ring in the flange. Those operations can be considered the preparatory phase for the real assembly of the next department. The assembly is not automated for all the 7 stations but some machines help the workers in doing the operations.

When a product exits the first assembly department, it is stored in a small FIFO buffer, whose size can contain the production of half a shift. Since that buffer is designed with only one entrance and one exit the FIFO policy must be strictly followed, namely only the first workpiece in the queue can be taken and brought in the second assembly department.

In that department there is a conveyor belt, which brings the parts from station to station, which are supplied with the needed components and where the planned operations are performed. In particular those operations consist in assembling all the remaining components, namely the shaft, the head, the ring, the junctions and so on.

When the parts exit the second assembly department, they are ready for the next two final areas of the line, i.e. the quality control and the painting department. In the quality control area some work benches are dedicated for each of the two types of product but the quality test is quite similar. First of all, there is the helium test. It means that some helium is pumped inside the pump and it is assured that the helium does not come out from the seals. Then the performances of the pump are tested with the fluid which they will elaborate in the real operational condition. If the testes are passed, the pumps are brought to the painting area, where again there are dedicated work benches.

Once they are painted, they are stored with the Kanban card still attached in the warehouse of the final products, until a costumer comes and takes them. When he does it, new Kanban cards are freed and they are taken in the Kanban board ready to be attached again for the production of a new product.

The mean production lead time, i.e. the average time from the withdrawal of a casing from the upstream buffer and the storage in the final product warehouse, is approximately 4 hours and the average work in progress in the plant can be estimated between 400 and 500 parts.

5.4 Customer

5.4.1 Available demand information

The customers communicate every month the indicative weekly demand for the next 6 months and every week the indicative daily demand for the next month. Then, they are let completely free to take the products from the final warehouse, also called supermarket, whenever they want and with whatever quantity and, as a result, the real demand profile is unknown. However, the indicative planned demand can provide precious in-

formation especially for what concerns closest periods, for instance, the next 2 weeks. Indeed, the company assumes that the daily real demand deviates from the scheduled one of ±20% at maximum. Moreover, it has been experienced that the daily scheduled demand is always in a range of ±40% the mean daily demand value. That means that the buffer must absorb and tackle two kind of demand variability. Firstly, the planned fluctuation because the scheduled demand is not constant on a daily base. Secondly, the demand uncertainty because the real demand can be different from the scheduled one.

5.4.2 Stockout and backlog

When a customer requires a product but it is not available, the demand is not lost but it will be satisfied as soon as possible. However, the company incurs some extra costs depending on which types of fines are defined in the contract which has been agreed with the costumers. Generally, if the blocking of the customer production system is not caused and the goods are again available in few days, the customer can come and take them with no additional expenses for the company. Another possibility is that, if no economical damage has been caused to the customer, the goods are delivered to the customer's plant within 24 hours at the expense of the company. On the contrary, if the customer is forced to block the production line, the company must pay him an amount of money which is defined by contract and is proportional to the amount of backlogged products and to the time needed to satisfy that demand. In practice, the real situation is really complex because there are many customers who have different contracts with the company. For the sake of simplicity, the model will consider an equivalent reasonable backlog cost and no stockout costs.

5.5 Kanban policy

Let us focus on how the Kanbans are used in the plant starting from the ones attached to the final products in the final warehouse. When some products are taken from it, some cards are freed and they go through different steps until they return to the final warehouse attached to a new finished product. The different steps are represented in the figure 5.6 and they are called RT_1, $RT_{lotCreation}$, RT_2, RT_3, RT_4, RT_5 and RT_6. In the figure 5.6 it is assumed that 4 Kanbans correspond to a production lot but in general each product can have its own lot size.

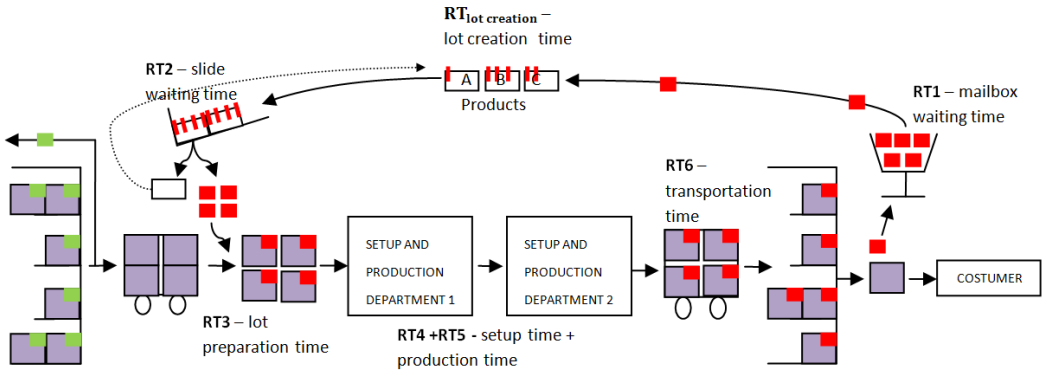

Figure 5.6: Kanban cycle

RT_1 represents the mean time that the Kanbans spend in the so called lettebox (Briefkasten). In fact, when they are detached from the final product, they are put in a box located nearby and only at the end of the day all the Kanbans, which are in the box, are taken at the same time to the next step. For that reason RT1 can be assumed equal to 12 hours, because assuming the arrival rate of freed Kanbans in the box constant, their waiting time is uniformly distributed between 0, if the they are freed at the end of the day, or 1 day, if they are freed at the beginning of the day. It means that the mean value is half a day. It would be not so economical and efficient to take them one by one or in small groups to the next step as soon as they are freed, although it would speed up their cycle time, because it would mean that extra workers and resources must be paid or an automated handling system must be designed to do it without having a great improvement of the system performances. Instead some extra Kanbans can be added in order to cover that delay.

$RT_{lotCreation}$ represents the mean waiting time of a Kanban that has just arrived in the area where the production lots are created. Indeed, the lot size is composed by more than one Kanban (it depends mainly on the available time to do setups) and each Kanban represents the production of more than one product. In fact, NPK is defined and corresponds to the number of products per Kanban and it is different for each product but a typical value in this plant is 50. The quantity associated to a card must be reasonably small on purpose because in that way a finite number of boxes of final product can be given to the costumers, whatever the required quantity is. Some small boxes are used to group the number of Kanbans needed to create a lot. They have a number of compartments equal to the number of cards needed to create a lot of a certain product and, as more and more Kanbans of that product come, they are inserted in those holes until the box is full. At this point the box with all the Kanbans is moved to the next step and placed on a slide. The lot formation is quite important to achieve a good performance of the system, because we would like to have a number of setups not too big, since they are considered as inefficiencies and take time, which would be instead available to produce more with the same amount of manpower. On the other hand, it must be also not too small because the line would be no more balanced and,

since a longer period elapses until a product is produced again, the risk of stockout for that product would increase. However, the average time spent in this section is very small compared to the other ones.

When the box is placed on slide, it must wait there for period equal to RT_2 until all boxes that arrived before are removed from the slide. Indeed, the slide follows a FIFO policy and it means that if one box was placed on it before another one, it will be removed also before. A Kanban waits here 8 hours on average

RT_3 includes the mean time needed for a Kanban to be removed from the slide and separated from the box, which is sent back to the area where the lots of Kanbans are created, and the time to prepare the new material for the production, namely the casings must be taken from the upstream warehouse, the Kanbans (the green ones in figure 5.6) must be detached and put in the mailbox of the upstream department, the new Kanbans (the red ones in picture 5.6) must be attached to the casings and they must be brought close to the beginning of the line ready for the production. A typical value for RT_3 is 1 hour.

RT_4 and RT_5 can be considered together and they represent the mean time, that a Kanban needs to go through all the production line, i.e. from the first department until the painting area. It corresponds to the sum of the production time, the setup time and the waiting times in some intermediate buffers and it is equal to more or less 4 hours.

When the finished product exits the production line, it must be stored in the downstream warehouse and for that reason we must consider also a RT_6, which considers the transportation time from the end of the production line to the warehouse along with the stocking time. Those two activities increases the loop time of Kanban of approximately half an hour.

5.5.1 Computation of the Kanban number

All those steps and their RTs must be considered in the computation of the Kanban number to insert in the production line. That computation is performed every 2 weeks and it also lasts 2 weeks along with the production plan decisions. The Kanban number computation is done for each product and for each productive phase (final assembly, all the upstream phases which produce components and theoretically even for the external suppliers). In this Thesis it is considered only the final assembly department, because it is where there is the most critical situation, since its warehouse absorbs and smooths the scheduled and not scheduled demand peaks of the costumer. However, the reasoning for the other ones is almost identical but since the production rate of the assembly department doesn't vary a lot and all the other lines are upstream with respect to it, there won't occur particular problems and the optimal number of kanbans is much lower).

The equation 5.1 represents the Bosch Kanban formula which is used to compute their number.

$$K = RE + LO + WI + SA \tag{5.1}$$

K is the total number of cards to insert in the system in the considered Kanban cycle

for a given product and it must be computed for each of them. Their total number will be computed by summing all the contributions.

It will now explained how each element of the Bosch Kanban formula is computed.

1. RE (replenishment time coverage) corresponds to the number of cards needed to cover the whole $RT_{loop} = RT_1 + RT_2 + RT_3 + RT_4 + RT_5 + RT_6$. The $RT_{lotCreation}$ will be considered separately in LO.

$$RE = \frac{RT_{loop}}{TT_{snr}NPK} \tag{5.2}$$

With:
$TT_snr = \frac{MTBO}{MOQ}$
$MTBO$=mean time between two orders of the considered product
MOQ=mean ordered quantity of the considered product
NPK = number of parts per Kanban

2. LO (lot size coverage) represents the number of Kanbans needed to cover the period required for the creation of a production lot.

$$LO = \frac{PPL}{NPK - 1} \tag{5.3}$$

PPL=number of parts that correspond to a production lot of the considered product

3. WI (withdrawal peak coverage) number of Kanbans required because the scheduled demand is not constant, as we have considered until now in the formulas ($meanDemand = \frac{1}{TT_snr}$), but it has some peaks.

$$WI = max(0, \frac{WA}{NPK} - RE - LO) \tag{5.4}$$

WA is the maximal aggregated scheduled demand with a length of RT_{loop} during the next two weeks.

For example, let us suppose that RT_{loop} = 3 days and the scheduled product demand PD is the one presented in figure 5.7

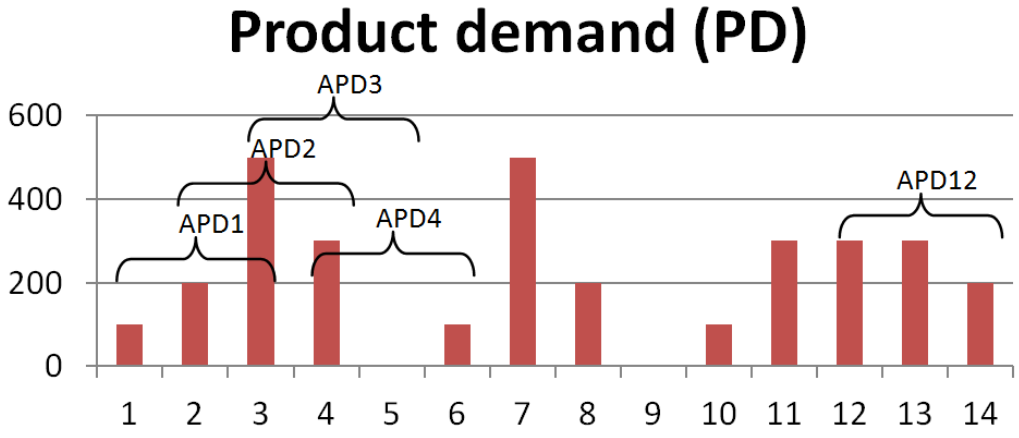

Figure 5.7: Product demand

The computed aggregated demand $APD = PD_i + PD_{i+1} + PD_{i+2}$ is represented in figure 5.8

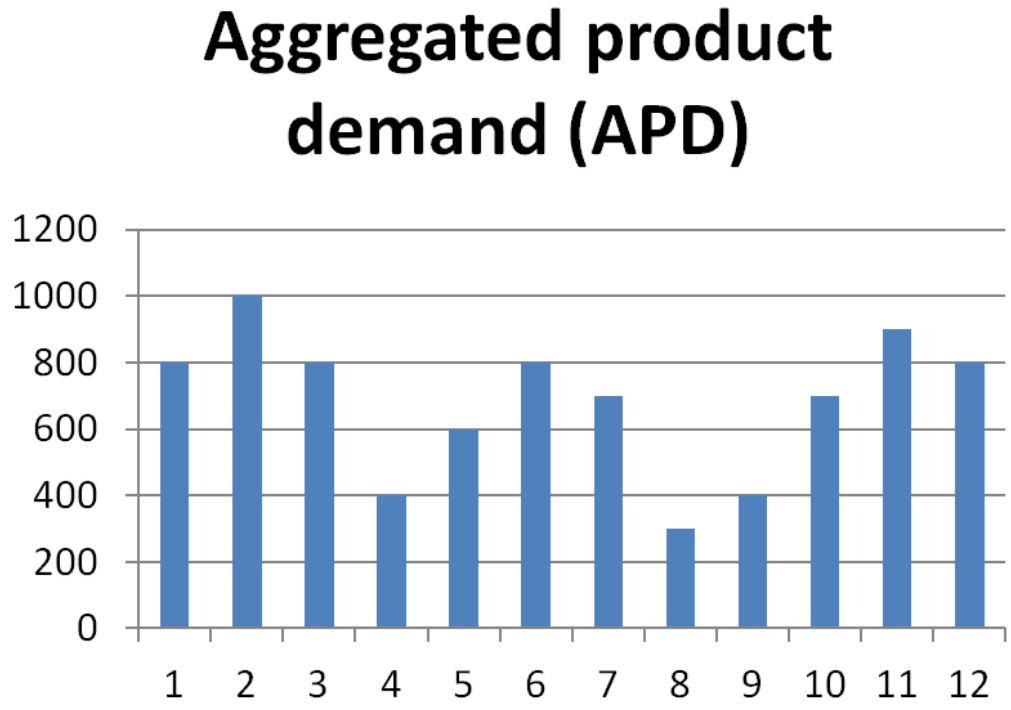

Figure 5.8: Aggregated product demand

In this example WA = 1000

4. SA (safety time coverage) is composed by three terms.
   - SA_1 refers to the Kanbans that take into account the unknown deviations or changes of the throughput and lead time of the process with respect to the forecast one. For example, if the real mean OEE value is 5% higher or lower than the value of the last 2 weeks, which is the value that is assumed for the next 2 weeks or if the OEE fluctuation is higher or lower than expected. Moreover, it can happen that the replenishment time is higher and it means that more Kanbans are required because the RE and WA would be bigger. Those deviations can happen for different reasons, i.e. unplanned plant stops, unplanned performance losses, etc.
   - SA_2 refers to the Kanbans which take into account the unknown deviations of the real demand with respect to the scheduled one. For example, if there are some random deviations between the scheduled and the real demand, it can happen that WA gets bigger and as a consequence more Kanbans can be necessary. In any case, the deviations can be of the two kinds : when the costumer takes a different quantity of material and when the costumer takes the same quantity but in a different time period. In the reality the two deviations are also mixed together.
   - SA_3 refers to Kanbans to have a greater safety and service level by taking into account some possible additional problems such as delays and errors in the information flow or even bigger unexpected deviations.

5.5.2 Considerations on the Kanban computation

In any case, all those three parameters are computed in a not rigorous way, especially for what concerns SA, without optimizing any cost objective function and by simply imposing some reasonable deviations and measuring how many Kanbans would be necessary in that case and, if the obtained number is lower, no more cards are used, otherwise their additional number for a certain unplanned phenomenon is equal to the difference. Another weak point of that approach is that it is not considered at the same time the effect of all possible deviations in the production system, all possible deviations in the demand and, optionally, some extra deviations but they are considered one by one in a separated way. The combined effect of two deviations can be much stronger than the sum of the deviations standalone.

For example, let us suppose that we are sure that the OEE mean and variance are estimated correctly but we suppose that the replenishment time can be 10% lower or higher that the expected value. The total number of Kanbans in the base scenario is already known, so its value must be computed for the other two scenarios and a higher value for RE and WI (and as a consequence for K) is expected when the replenishment time is increased by 10% because RE has RT_{loop} on the numerator and WI depends on WA, which increases when RT_{loop} increases. For those same reasons a lower value for

K is expected when the replenishment time is decreased by 10%. In that situation it should be considered the base scenario and the scenario with +10% replenishment time, compute K for the latter and find $SA1 = K_{+10\%scenario} - K_{baseScenario}$.

Let us consider an example where the daily demand of a certain product for the following two weeks is known with an uncertainty of ±10%. In that case, we must consider three scenarios, i.e the base scenario and the two scenarios with +10% and −10% respectively, and compute the value of the suitable Kanban number for all of them. The SA_2 can be computed with the formula 5.5.

$$SA2 = max(0, K_{-10\%Scenario} - K_{baseScenario}, K_{+10\%Scenario} - K_{baseScenario}) \tag{5.5}$$

With those approximated formulas the computed number of Kanbans results as a matter of fact always so big, that the production is almost never blocked due to the absence of Kanban cards as it can be seen in the historical production data. Even if no cards are available, some special temporary red Kanbans are added (and then removed as soon as they are detached from the final products) and make the production continue regularly, instead of blocking it and affect badly the OEE of the day. It is reasonable because, if we would want to limit the number of cards, it is because we would like that the production is blocked when the inventory is almost full but it would mean that when the production line is blocked, there are some workers that are paid but they are not allowed to work. For that reason and since the manpower in this production context costs a lot, it is better from an economical point of view to adjust the production quantities with a reaction policy which adds and removes shifts and/or employees than suddenly block the production line for the absence of Kanbans.

5.5.3 Usage of the Kanban policy

However, the Kanbans play an important role. First of all, the leveling policy works well when the real demand is predictable with a good approximation taking into account the scheduled demand provided by the customers. The opposite situation occurs when there is the so called "fire fighting" which means that the scheduled demand has become quite unreliable and a different product mix will be required day by day. In that case the advantage of planning the production along with the leveling in advance decreases because the information is less reliable and the usefulness of the Kanban cards increases because the employees at the beginning of the production line have a quick feedback about which products have been taken by the customers by considering the free Kanbans.

Nevertheless, the production plan is scheduled in any case in advance because the "fire fighting" is neither predictable nor frequent. Then, if less or no Kanbans are available to respect the production plan, it means that probably the demand quantities and/or mix was different than expected. When it happens, we must define a new plan for the next days and the so called "go & see" should be done, which means that some employees or managers must go to the final warehouse and find out which Renner products have a dangerously low inventory level. After that, the production plan of the next days is

changed by taking into account that information. Instead of the "go & see" procedure it is easier and quicker to check the available Kanbans at the beginning of the line and to identify the products with the higher number of available cards. A new production plan for the next days is defined by producing more those products instead of the ones, whose inventory level is already high and by taking into account all the considerations and constraints of the production planning, which will be discussed in the following section. However, it is generally preferred to change only the production mix and not the scheduled production quantities, which are not changed to keep the same planned workload in the department.

5.6 Production policy

As previously discussed, a rigorous Kanban-policy, namely that the production is not scheduled in advance but it is authorized only by available Kanbans at the beginning of the line, works well only when the demand has small fluctuations. Indeed, if the demand is stable, the production line will produce more or less the same quantity each day and it means that the line and the workload of the workers will be well balanced and the resources will be used always efficiently.

By contrast, if the demand has big fluctuations, it can happen that in some days there are few available Kanbans because only few costumers need parts and, as a consequence, the assembly line will produce only a small amount of products and there will be long idle times. Then, it is possible that in the following days the demand is very high because a lot of costumers are running out of components in the same days. As a result, the system will not be able to produce the required quantities, since it is normally designed to satisfy only the mean demand to limit the investment costs, and the buffer level will dangerously decrease risking the stockout. In order to avoid it, the company is necessary to keep high inventory levels for the Exoten products for a long time without taking advantages on it and incurring additional inventory costs. Furthermore, the production resources would be not efficiently used due to the aforementioned idle times.

Secondly, it could not be applied the EPEI (every part every period) policy, which consists in producing different kinds of products and doing a reasonably big daily number of setups to balance the production line. Indeed, since the line is balanced considering an equivalent product, which incorporates the mean production times of all the others, it can happen that, while producing the product A, the first department is faster than the second one and the intermediate FIFO buffer will tend to get full and, as a consequence, the first department will be forced to slow down and vice versa while producing the product B. If a lot of setups are planned, the line will be more balanced, the FIFO buffer will absorb the difference in the production rates between them and, as a result, the system will perform better.

5.6.1 Available time to produce computation

For those reasons the production managers along with some other employees (maintenance employees, some workers, and so on) have a scheduled biweekly meeting, where they decide what will be produced in the following 2 weeks. Those decisions are made with the following rigorous method. In order to better understand, it is a good idea to keep in mind the diagram in figure 5.9 which presents the division method for the total daily production time.

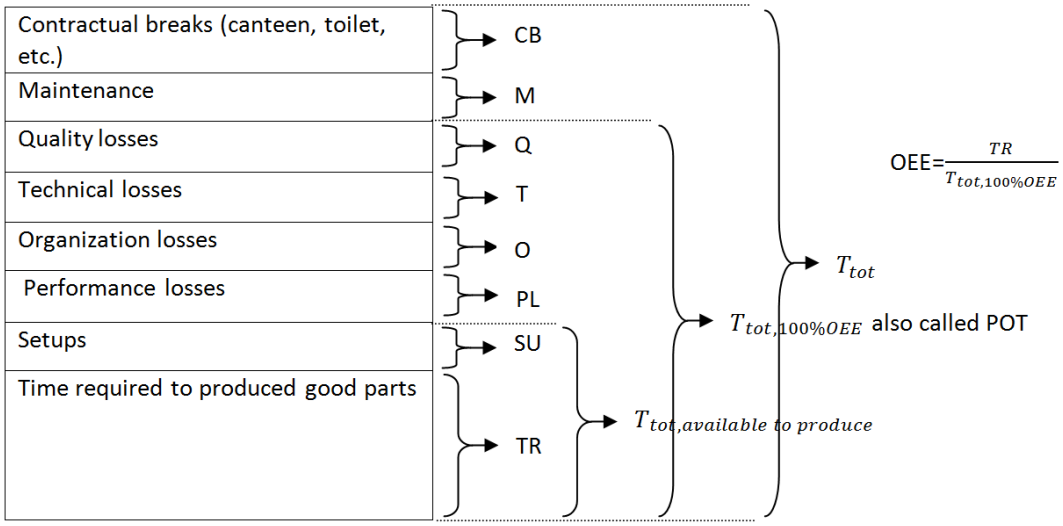

Figure 5.9: Daily time division

The maximum time available is always the same, namely 8 hours or 480 minutes per shift. It means that for a day with 3 shifts, $T_{tot} = 3 \cdot 480min = 1440min$.

Then the contractual breaks, which represent the scheduled losses and include mainly the unproductive time of the shift changes and physiological breaks, and the time required for the TPM (total productive maintenance), which are equal to 195 minutes/day on Tuesday and 150 minutes/day on the other days of the week, must be subtracted. Indeed, there is a rigorous plan of scheduled preventive maintenance, which consists in adding some lubricant to the machines and substituting some components regularly. Let us consider a Monday for the example. $T_{tot,100\%OEE} = T_{tot}$–$150min = 1290min$ Now we must subtract the not scheduled losses, which can be grouped in 4 classes:

- Quality losses which includes the equivalent lost time when some parts are found defective or when a portion of the line is blocked to solve those kind of problem
- Technical losses which includes the equivalent lost time when some stations of the line are failed or can only work at lower rate, some blackouts occur and so on
- Organization losses which includes the equivalent lost time due to strikes, assemblies, lack of components, absence of some workers and so on

- Performance losses which represent the equivalent lost time when the workers produce at a lower rate than the nominal one

Since they are not planned , their percentages with respect to T_{tot} are not known in advance but they are currently assumed equal to the ones of the previous weeks. For our example some reasonable values are assumed.

For the last 2 weeks :
$\frac{Q_{last2weeks}}{T_{tot,100\%OEE}} = 0.37\%$
$\frac{T_{last2weeks}}{T_{tot,100\%OEE}} = 2.46\%$
$\frac{O_{last2weeks}}{T_{tot,100\%OEE}} = 1.68\%$
$\frac{PL_{last2weeks}}{T_{tot,100\%OEE}} = 4.29\%$

Next 2 weeks :
$Q_{next2weeks} = 0.37\% \cdot T_{tot,100\%OEE} = 4.8min$
$T_{next2weeks} = 2.46\% \cdot T_{tot,100\%OEE} = 31.7min$
$O_{next2weeks} = 1.68\% \cdot T_{tot,100\%OEE} = 21.7min$
$PL_{next2weeks} = 4.29\% \cdot T_{tot,100\%OEE} = 55.3min$

It is now possible to compute $T_{tot,availableToProduce}$.

$$T_{tot,availableToProduce} = T_{tot,100\%OEE} - Q - T - O - PL = 1176.5min \tag{5.6}$$

$T_{tot,availableToProduce}$ represents the available time both to produce good parts and to make setups. The managers must then decide how to divide it between SU and TR. As already discussed, there is a trade-off between having a big number of setups, which assures to have a balanced production line, and a small number, which implies less time losses due to setups. The solution is currently found without a rigorous method.

5.6.2 Reaction policy

The reaction policy is basically performed by adjusting the number weekly planned shifts according to the buffer level. The parameters can be only decided every 2 weeks and the remain frozen for that period.

5.6.2.1 Escalation ranges

In this production plant different ranges for the monitored index, i.e. the inventory level, are defined and each of them is associated to a precise level of alert. It means that a normal range is defined and it represents the most desirable one for the monitored index because both the risk of stockout and the inventory costs are low and usually, since the plant is designed to be cost-efficient in the normal operation condition, also the operational costs will be low with respect to situations where the plant is forced to produce less or more than the nominal rate. After that, three alert ranges are defined for situations where the inventory level exits the normal range and, as a result, there are basically four index ranges.

1. Range 1. When the buffer level is dangerously low, the plant tries to produce more than required in the next 2 weeks because the risk of backlog is becoming higher and higher

2. Range 2. Normal inventory range. If the inventory level in this range, the plant tries to produce more or less the same quantity as required to maintain the buffer in the same range for the next 2 weeks

3. Range 3. When the inventory level is slightly big, the plant tries to produce a bit less than the required quantity for the next 2 weeks.

4. Range 4. When the inventory level is dangerously high, the plant tries to produce an even lower quantity than the required one for the next 2 weeks

The buffer fluctuations, associated to high stockout and blocking risks, could be also limited without the 3^{rd} range because there already exists the 4^{th} range which tries to decrease the buffer level when it is big. However, it can improve the cost-efficiency of the system, as demonstrated in the final chapter, because, although the 4^{th} range provides a stronger reaction, it generally causes a lower cost-efficiency. Indeed, if the production quantities must be strongly lower (or greater) than the required quantity, which is generally similar to the plant nominal production quantity, the system will be forced to work far from the nominal operational conditions and, as a result, with a low cost-efficiency.

5.6.2.2 Adjustable parameters - number of shifts

As already said, the production quantities for the next 2 weeks are currently chosen according to the current buffer level during the biweekly meeting. Then, a production plan for that period is created by choosing the number of shifts and the production mix for each day. The first one can not be changed until the next biweekly meeting, while the second one can be slightly modified if the "fire fighting" situation occurs.

The nominal production rate of the plant can also be modified by changing the amount the manpower but it parameter is changed very seldom and only for these three reasons. Firstly, to slightly and quickly adjust the production rate when, for instance, some urgent orders are added or the stockout of a product is currently risked. Secondly, when other lines require more manpower and it is economical to move it away for a limited period. Thirdly, when some training courses are planned. In any case, in those three cases the manpower is not used as adjustable parameter to control the inventory level and, as a result, this parameter is considered in the model only as additional production variability and not as an adjustable parameter.

As a consequence, the system reconfiguration is done by changing the number of scheduled shifts in the week which identifies the expected production quantities. By considering this adjustable parameter, 4 possible number of weekly shifts can be used.

1. 15 shifts per week, which is the basic shift plan and it corresponds to produce the first 5 days of the week with 3 shifts per day.

2. 14 shifts per week, when only 1 shift is removed and it is generally removed the last shift on Friday or the first shift on Monday since it is preferable to maintain the production continuous along the week

3. 12 shifts per week, when an entire day of production is removed from the basic shift plan. The removed day is usually Friday or Monday

4. 16 shifts per week, when an extra shift is added on Saturday and precisely immediately after the last shift on Friday

It can also happen that, for example, when some shifts must be removed, the production is not maintained continuous but it is generally preferable to do so because the plant restart has also a cost.

By considering the current inventory level during the biweekly meeting, the shift number for the first week is chosen. Then, the inventory level at the beginning of the second week is estimated and the shift number for the second week is also chosen. However, this estimate is made without any rigorous method and, as a result, the estimate can be very poor.

5.6.2.3 Daily schedule and leveling

Once the production quantities are defined, it must be decided which products must be produced during each day and how many times. After that, it must be decided the daily scheduled production sequence. In order to do it, the managers must take into account the inventory level of each Renner because, if its inventory level is dangerously low, that Renner must be produced, even if it is not included in the scheduled demand of the next 2 weeks, to cover an unexpected demand. Its scheduled production quantity will be not very big but enough to cover an unexpected demand. There are not available tools to rigorously estimate the magnitude of a possible unexpected demand but a reasonable value can be hypothesized.

The Exoten are managed with a push-technique and, since there are no already finished Exoten in the warehouse, it must be assured that they will be available when the costumer will come and take them to avoid backlog costs. For what concerns the Renner, they can be shifted without many constraints in the considered period because there are already some products in the warehouse ready to be taken, so they can also be produced when the costumer has already taken them.

Since the scheduled quantities are identified by the current inventory level, they are in general different from the scheduled demand quantity. For example, if the final warehouse is already full and, as a result, it is planned to produce with only 12000 parts per week but the scheduled demand to be leveled consists of 15000 parts per week, the quantities to be leveled must be multiplied by $\frac{12000}{15000}$ to obtain the rescaled leveling quantities and to meet the defined production volumes. A similar reasoning can done if the decided production quantities are bigger than the scheduled demand.

Taking those considerations into account, the leveling must be performed. First of all, the scheduled demand day by day must be considered. It must be rearranged to have a

more regular production and a more balanced production line ([CS00]). That procedure is called leveling (Nivellierung). Of course, there are plenty of possible production plans and to identify them and the equivalent ones, the parameter EPEI (every part every interval) is defined. It represents the time interval, where all the Renner products are produced. For example, EPEI=1 means that the scheduled production has been rearranged in such a way that all Renners are produced at least every day. Moreover, EPEI=0.5 means that all the Renners are produced every half a day and if EPEI=2 every 2 days. On one hand, it would be better to have a higher EPEI because it means that less time is lost in doing setups and, as a consequence, the setup losses will decrease and the plant can produce theoretically more with the same amount of manpower. On the other hand, it is also an advantage to produce each Renner continuously (for example every shift or every day) for two main reasons. First of all, the line can be more balanced. It can happen that the first department is faster than the other in producing the product A and the second is faster with product B. If the plant does not switch between A and B often, the consequence would be that, when A is produced, the FIFO buffer between the departments tends to be full and vice versa and that situation slows down the production. The second reason is that, if a certain product is produced only every two weeks, it will be easier to have a stockout for it.

The figure 5.10 helps to understand how exactly the production plan changes according to the EPEI.

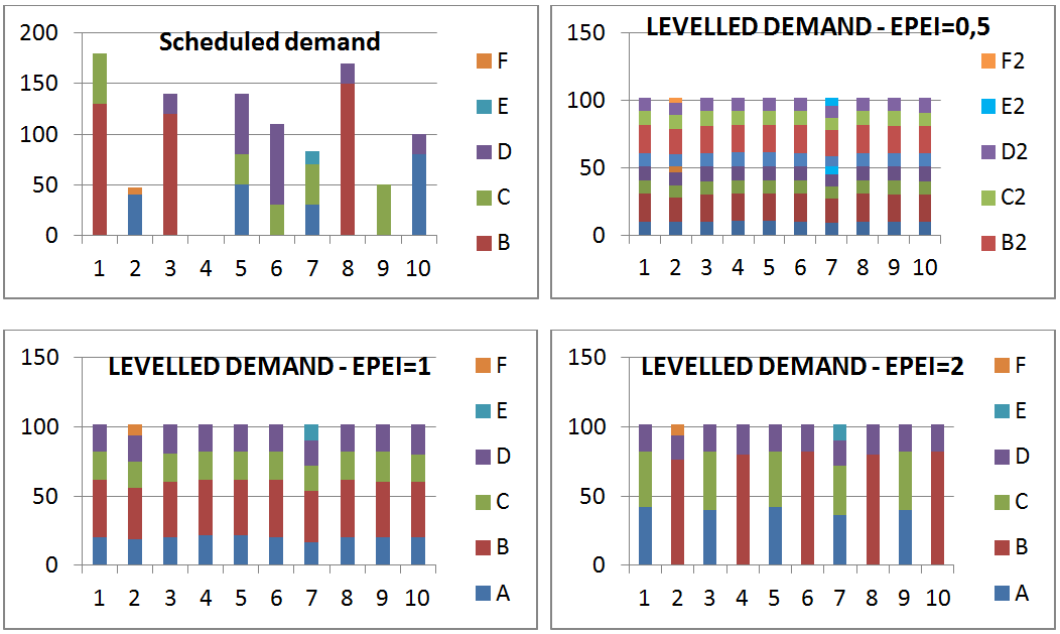

Figure 5.10: Leveling for different EPEI values

The optimal value for EPEI is also a function of the number of different products and the scheduled quantities of each one. For example, if the number of different products

gets bigger, the optimal EPEI will also increase because the setup frequency would increase and cause too many OEE losses.

The production sequence, i.e. which Renner products will be produced before others in the time period defined by EPEI, is generally not important because there is usually an already produced quantity for each product in the final warehouse enough to satisfied its demand of more than one EPEI period. If it is not true it must be decided to produce some chosen product, whose buffer level is low, at the beginning of the EPEI period.

During the EPEI choice the only constraint which we must take into account is that for every single day of the next two weeks it must hold:

$$N_{partsToBeProduced} \cdot Takt \leq T_{tot,availableToProduce} - N_{setups} \cdot T_{setup} \tag{5.7}$$

It means that $TR + N_{setup} * T_{setup} \leq T_{tot,availableToProduce}$
$N_{setup} * T_{setup} \leq T_{tot,availableToProduce} - TR = SU$
In other words, once the quantities to produce in one day are defined, we must compute the daily remaining time for the setups SU and their number must be $N_{setup} \leq \frac{SU}{T_{setup}}$.

5.6.2.4 Considerations on the current reaction policy

The weak points of this method are several. First of all, the real costs which the choice implies are not explicitly taken into account and the reaction levels are not computed with the cost optimization of the escalation levels depending on their position. Secondly, they are not updated, if some characteristics of the production plan or of the demand or the cost coefficients change. It is of course reasonable because, for instance, if the production variability and the demand uncertainties decrease, the backlog risks will be more predictable and the system will react only if the inventory level is very low and, as a consequence, will use more the cost-efficient production plans.

5.7 Anticipated replanning

As it was already said, the production plan is defined during the biweekly meeting for the next 2 weeks. However, if the characteristics of the plant or of the demand drastically change during that period and, as a result, the escalation levels are no more optimized for that production context, it can be profitable to reschedule the production quantities by organizing an unscheduled meeting. It is important to understand that an extra meeting has a cost, which must be compared to the economical advantage of the new optimization. The costs to be considered are mainly the costs associated to the fact that the managers must participate to the extra meeting instead of doing some other tasks as, for instance, help the production line, which implies some extra costs or extra inefficiencies. Since it is hard to quantify those extra costs, the decision must be taken without a rigorous method. For example, if the buffer is emptying due to an unexpected serious failure on the line and it is probable that a backlog situation will occur with the current production plan before the next meeting, it is probably better to plan an extraordinary meeting to increase the number of shifts at the end of the week.

Fortunately, it happens very seldom that the plant and demand characteristics change so drastically that it is better to plan an extra meeting and, if those drastic changes happen few days before the scheduled meeting, it is not decided to plan the extra meeting, because the problem would be tackled in few days.

6 Bosch case modeling

In this chapter it is explained how the presented production context composed by the Bosch plant, its reaction policy, the buffer and the demand behavior can be modeled using the methodology presented in chapter 4. Then, it is shown how the costs of the current policy can be evaluated and how much money could be saved if the optimal policy suggested by the optimization algorithm would be used instead of the current one.

6.1 Upstream machine modeling

In this section the modeling of the upstream machine is presented. The starting point are the historical data of the past year (2013), which include the many pieces of information like the daily available time to produce, the daily produced parts, the amount of daily lost time of each day, etc. Those data are used to estimate production rate distribution and to investigate whether it is influenced by the number of planned shifts in the day or by the day of the week. The results are then used to estimate the daily production expected value and variance for each plan and the corresponding Markov chain. Finally, it is shown how the delays in the reconfiguration can be modeled.

The upstream machine of the multi-threshold model represents the final assembly department of the production plant. It can produce with 4 weekly production plans which are characterized by the number of weekly shifts. A different number of shifts implies that the plant works for a different amount of time every week. On the other hand, the customer can require products only form Monday to Friday. It means that the two machines do not work always simultaneously on a daily basis. In particular, they do it when the planned shifts are 15 but they do not otherwise. Since the corresponding approximation is small, it is reasonable to model that both stages work always simultaneously, as if also the production plant worked always from Monday to Friday but the production rate was dependent on the number of shifts. As a result, the system is modeled on a weekly basis, i.e. that only the weekly production distribution of each plan is considered and not the daily ones.

6.1.1 Available data

The production distributions are estimate by considering the historical data which include the real daily production of an entire year. Since the main goal of this section is to model the production plant in the most common situation and since the systems characteristics can change during the year, the days where something exceptional and

not related to the normal production variability happens must be identified and not considered in the computations. Particularly, it is decided to exclude the days belonging of those three groups:

1. Group 1. Days whose Takt is different from 0.417 $\frac{min}{part}$. Normally, when the Takt is greater than 0.417 $\frac{min}{part}$, it means that some manpower has been moved to other lines. By contrast, when the Takt is lower than 0.417 $\frac{min}{part}$, it means that some additional manpower has been taken from other lines. However, in the historical data this kind of relation is not always true and it is likely that there were some errors during the data recording and/or typing. Those unusual days are few compared to the total number of available days in the historical data (10 out of 207) and, as a consequence, they can be deleted without affecting the power of the inference on the μ_u. The days with a different Takt are included in the following table.

Day	$Takt$
$17^{th} January$	0.578
$21^{rt} January$	0.502
$12^{th} February$	0.431
$8^{th} March$	0.344
$9^{th} July$	0.402
$6^{th} August$	0.431
$10^{th} September$	0.486
$8^{th} October$	0.431
$7^{th} December$	0.332
$10^{th} December$	0.402

2. Group 2. When some extraordinary maintenance is planned. Since it is a planned loss, it would not be correct to consider it as a random failure. The historical data show that an extra maintenance was planned only on the 31^{st} of January and it stopped the plant for 180 minutes.

3. Group 3. Unusual days where the OEE exceptionally decreased due to expected problems related, for example, to a certain period of the year or some expected events. The days with those characteristics are shown in the following table.

Day	OEE
$4^{th} January$	69.90%
$5^{th} August$	68.35%
$6^{th} August$	84.52%
$7^{th} August$	77.52%
$8^{th} August$	80.62%
$9^{th} August$	64.47%
$26^{th} August$	75.71%
$27^{th} August$	74.73%
$28^{th} August$	79.59%
$29^{th} August$	87.27%
$11^{th} December$	76.10%
$27^{th} December$	70.03%

The OEE of all those excluded days is much lower than the target value (85%) which is chosen by the company as a minimum minimum value to be reached or the mean OEE value (approximately 89%) of the rest of the year. Firstly, it is decided to exclude all the days belonging to the two weeks of August where mean OEE is very low reasonably due to the fact that lots of workers requested vacation days in those periods and, as a consequence, the amount of manpower working on the line was lower (as it can be seen in the historical data). Secondly, the $4^{th} January$ and the $27^{th} December$ are excluded as well because there was an exceptional lack of manpower which also due to a simultaneous request for vacation days. On the contrary, it is decided to consider the small absence of manpower as unexpected because not related to particular period of the year. Thirdly, the $11^{th} December$ is also excluded because there was an organizational loss of 4 hours and a half due to a strike. Since it is usually known during the meeting if strikes are planned in the next two weeks, this plant idle time was not unexpected. It must be noticed that in all those cases the Takt (i.e. the planned cycle time) of the day is not changed because the lack of manpower is not due to internal needs but it is considered as an organizational loss.

If those days were considered, the normal variance of the production volumes would be overestimated and the normal mean OEE would be underestimated leading to a wrong cost evaluation. It is preferable not to consider them because in a normal period of the year those situations never occur.

In order to identify the first and the second group of weeks to be excluded, the number of planned parts to produce can be checked. That number make us able to detect if the managers decided to produced less or more than the normal production quantity, which can be computed by multiplying the POT of that day by $\frac{1}{Takt}$, because some extra losses were already known during the biweekly meeting.

6.1.2 Rate distribution

Once all the usual days are selected, it is possible to divide them in three groups in relation to the number of shifts included in each day (1,2 or 3 shifts) and it is possible to compute mean production rate for all the samples with the equation 6.1

$$\mu = \frac{produced\ Parts}{available\ Time} \tag{6.1}$$

It is decided to test the production rate, for instance instead of the daily production volumes, because, even if the number of daily shifts in each group is the same, the time available to produce can be different due mainly to the extra maintenance on Tuesday. As a result, if, for example, two days with three shifts are considered and the first one has $POT = 1290min$ and the second one $POT = 1245min$, it is reasonable to think that the expected production quantity is different but the expected production rate is the same. Moreover, it is considered the daily production and not the weekly production because, in the second case, if one day is evaluated as unusual, all the other days must be also excluded. As a consequence, with the used procedure it is possible to exploit a greater number of data which implies more power in the statistical tests.

At this point we want to test if the mean daily production rate is influenced by the number of planned shifts in that day. Basically, a one-way Anova with 3 levels must be performed.

The 3 levels are:

1. Production rate of days with 1 shift

2. Production rate of days with 2 shifts

3. Production rate of days with 3 shifts

On the first level there are 3 samples, on the second 6 and on the third 195. It is immediately notable that the power of this test will be low especially if the expected value of the first level is different but anyway some useful consideration can be done on the results of the Anova test.

In order to decide which kind of ANOVA (traditional or parametric) is better to be used to test $\hat{\mu}$, its distribution shape can be qualitatively investigated. In particular, it can be estimated by considering only the 195 samples of the first group (since it is not known whether the three groups of μ come from the same population, they can not be mixed). The graph in the figure 6.1 is determined.

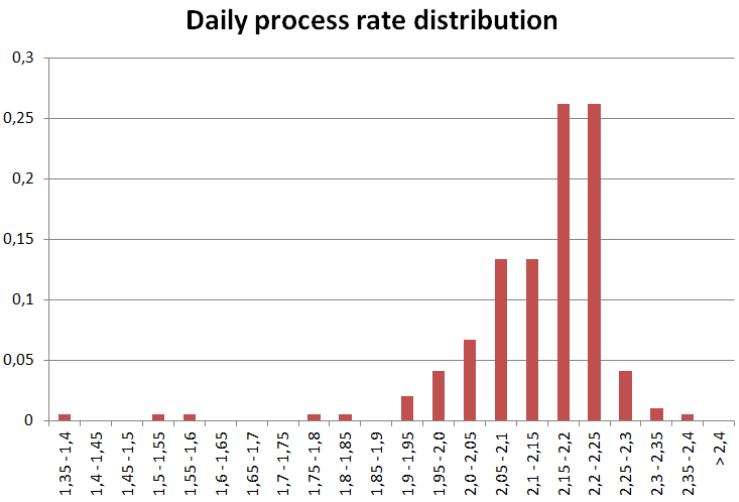

Figure 6.1: Rate distribution

The distribution looks absolutely not gaussian but there is an upper limit which coincides more or less with the nominal production value which the company associates to the situation when the OEE=100%. This value is always not reached because there are always at least some small OEE losses during a whole day and, as a result, the mode of the distribution is between 2.15 and 2.25 which correspond to an OEE between $2.15\frac{part}{min} \cdot 0.417\frac{min}{part} = 89.7\%$ and $2.25\frac{part}{min} \cdot 0.417\frac{min}{part} = 93.8\%$. Furthermore, there is a heavy left-tail which represents the fact that some exceptional problem can occur and the daily OEE can be exceptionally low.

In order to support our decision with a qualitative tool the normality test can be performed on the data (Figure 6.2), which underlines that the data are clearly not normally distributed since the p-value is by far smaller than 5%.

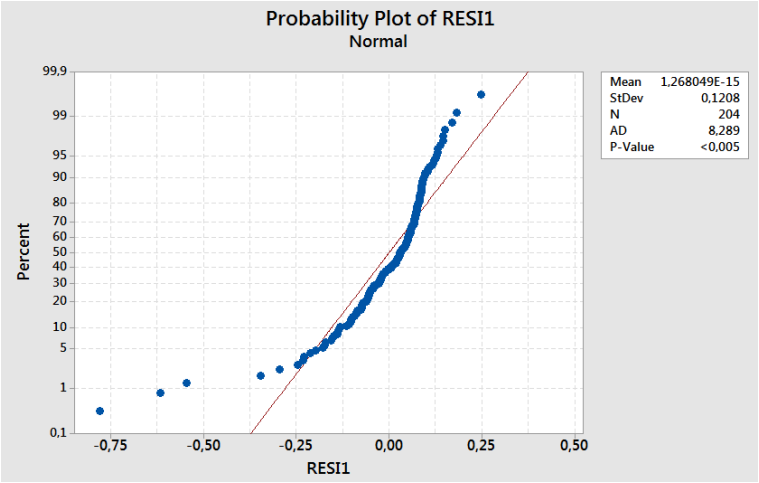

Figure 6.2: Normality test

6.1.2.1 non-parametric ANOVA on the number of daily shifts

It is now used the Kruskal-Wallis Anova which is the non-parametric equivalent of the one-way traditional Anova and it only assumes that the observations in each group come from populations with the same distribution shape, i.e. same variance and same distribution type.

First of all, the boxplot of the data (Figure 6.3) and the scatterplot (Figure 6.4) are created and some outliers are noticed on the third level but only on one side. They are a consequence of the non-normality of the data.

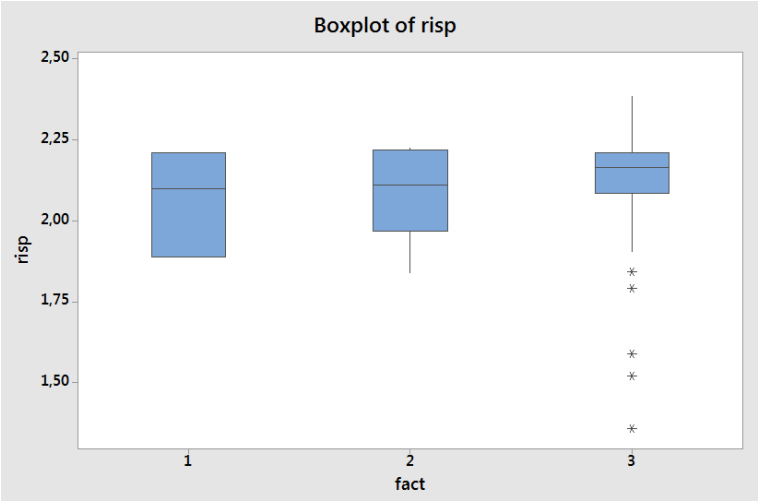

Figure 6.3: Data boxplot

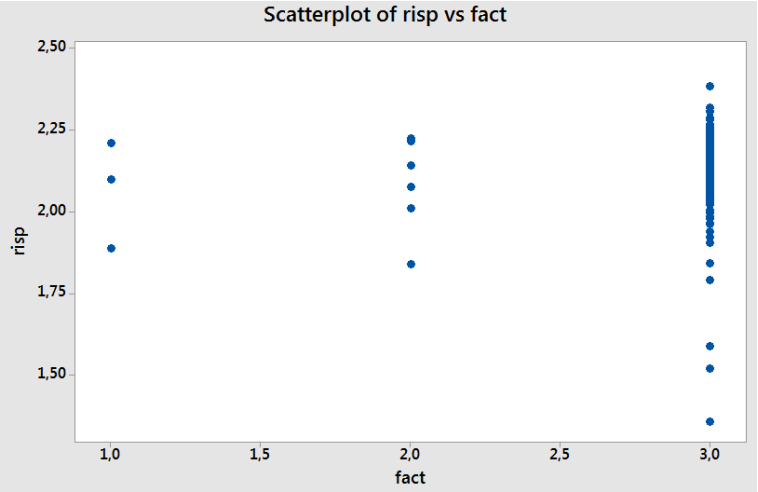

Figure 6.4: Data scatterplot

The interval plot (Figure 6.5)indicates that there is a possible trend of the response but, without a quantitative test, it is not possible to draw any conclusions.

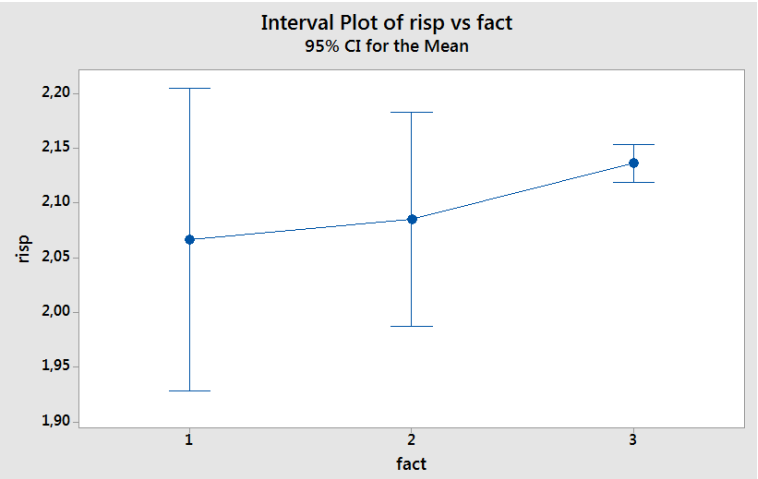

Figure 6.5: Data interval plot

The Anova table (Figure 6.6) shows that the null hypothesis of equal expected values can not be rejected with p-value=53.1%. For that reason it can be assumed that the null hypothesis holds if the assumptions are verified.

```
Kruskal-Wallis Test on risp

fact       N  Median  Ave Rank      Z
1          3   2,100      76,2  -0,78
2          6   2,109      83,6  -0,80
3        195   2,165     103,5   1,11
Overall  204             102,5

H = 1,27  DF = 2  P = 0,531
H = 1,27  DF = 2  P = 0,531  (adjusted for ties)
```

Figure 6.6: Kruskal-Wallis ANOVA

It is now needed to check the assumption of the non-parametric ANOVA test.

The test for equal variance (Figure 6.7) shows that the hypothesis of equality of variances can not be rejected with p-value=62.9% according to the Levene test.

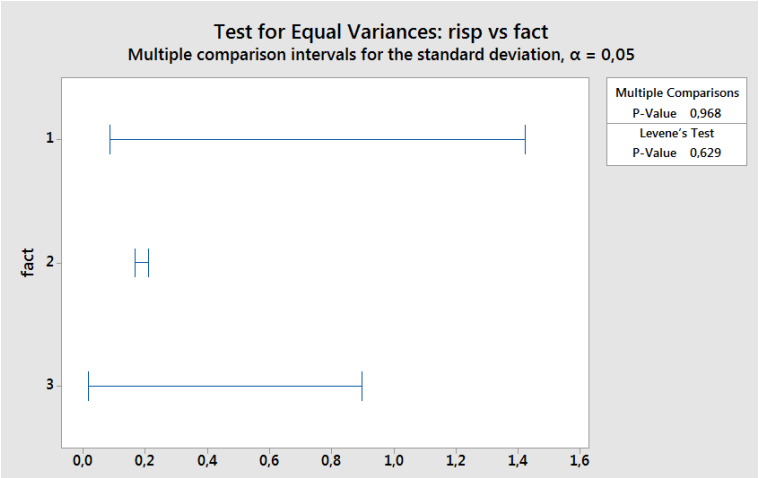

Figure 6.7: Test for equal variances

For what concerns the distribution type, there are not enough data to test it and it is only possible to assume that the three populations have also the same distribution type.

Finally, the autocorrelation test can not be performed because the order of the data is known but some data have been deleted because they did not belong to the population which must has been tested.

6.1.2.2 non-parametric Anova on the week days

It is now tested whether the expected value of the production rate varies in the different days of the week. The data are divided in 6 groups (From Monday to Saturday).

Day	$N_{samples}$
$Monday$	42
$Tuesday$	38
$Wednesday$	44
$Thursday$	43
$Friday$	34
$Saturday$	3

Again the power of test in detecting small differences on Saturday is low but for the other levels there is a number samples big enough to make significant test. The non-parametric Anova is used since it is now known that the data are not normally distributed.

From the boxplot,the scatterplot and the interval plot (Figure 6.8and 6.9 and 6.10)it seems that no differences between the levels exist.

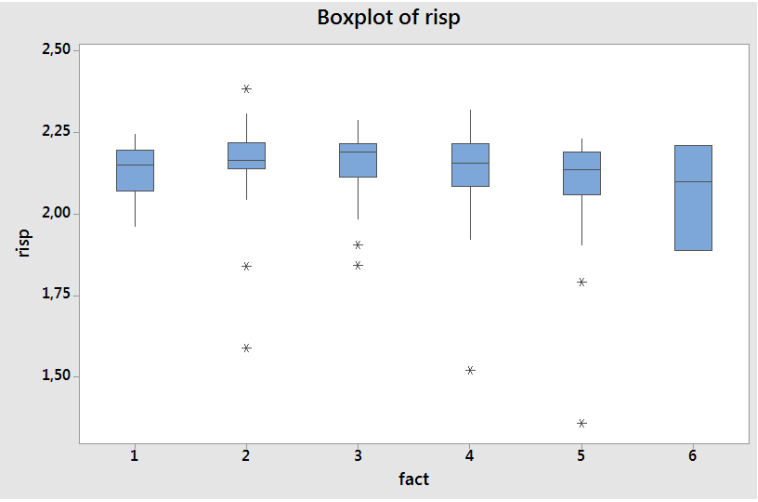

Figure 6.8: Boxplot

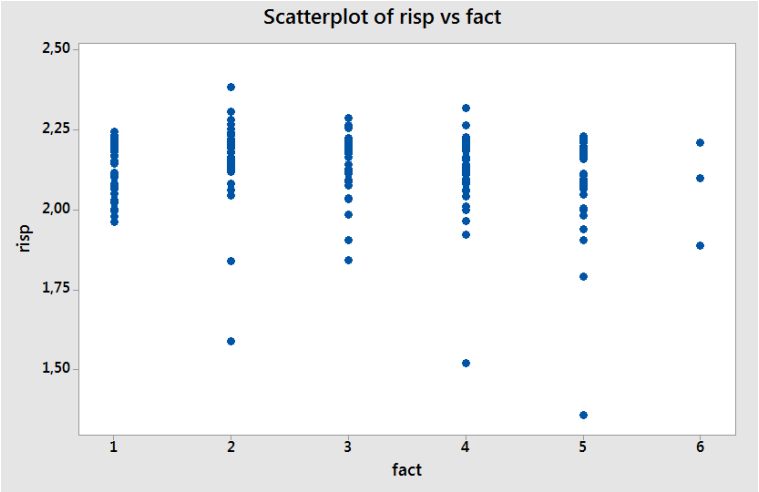

Figure 6.9: Scatterplot

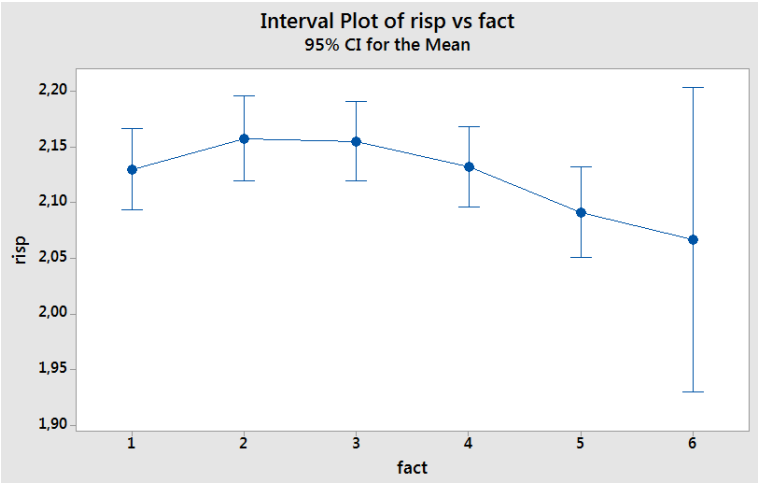

Figure 6.10: Interval plot

The Anova table (Figure 6.11) shows that the null hypothesis of equal expected value between the levels can not be rejected with α=5%, if the assumptions will be verified.

Test for Equal Variances: risp vs fact

Kruskal-Wallis Test: risp versus fact

```
Kruskal-Wallis Test on risp

fact       N  Median  Ave Rank      Z
1         42   2,150      91,4  -1,37
2         38   2,165     116,4   1,61
3         44   2,191     114,8   1,55
4         43   2,156     104,6   0,26
5         34   2,136      84,6  -1,94
6          3   2,100      76,2  -0,78
Overall  204              102,5

H = 9,27  DF = 5  P = 0,099
H = 9,27  DF = 5  P = 0,099  (adjusted for ties)
```

Figure 6.11: Anova table

Once again, it is assumed that the distribution types are the same and only the variances are tested with the Levene's test as shown in figure 6.12.

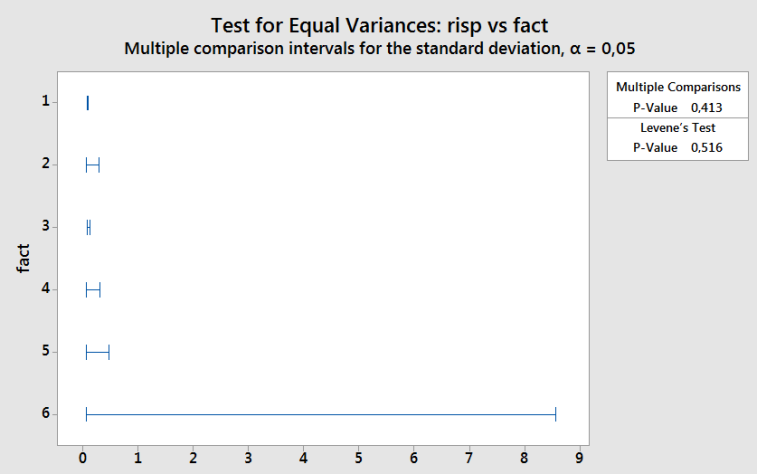

Figure 6.12: Test for equal variances

The hypothesis on variance equality can not be rejected with α=5%. The conclusion is that there is no evidence against the fact that the data of the three groups come from the same population.

6.1.3 Production plan characterization

Since it has been concluded that the production rate distribution does not vary in the different days of the week and there is no practical evidence that it varies during the day (morning, afternoon, evening and night), it is assumed that the production behavior is always the same. As a result, the equation 6.2 can be written. $\hat{X}_t$ corresponds to the production quantity produced in the t^{th} unit period of the overall considered time period ΔT.

$$\hat{X}_{\Delta T} = \sum_{t=1}^{\Delta T} \hat{X}_t \tag{6.2}$$

Since the production rate never varies, all the summed $\hat{X}_t$ are equal and it is possible to is possible to write:

$$E(\hat{X}_{\Delta T}) = E(\hat{X}_t) \cdot \Delta T \tag{6.3}$$

$$VAR(\hat{X}_{\Delta T}) = VAR(\hat{X}_t) \cdot \Delta T \tag{6.4}$$

Since $\hat{\mu}_{\Delta T} = \frac{\hat{X}_{\Delta T}}{\Delta T}$, it is easy to show that the expected value of $\hat{\mu}_{\Delta T}$ does not depend on the length of ΔT because the mean production quantity is proportional to it (equation 6.5).

$$E(\hat{\mu}_{\Delta T}) = E(\frac{\hat{X}_{\Delta T}}{\Delta T}) = \frac{E(\hat{X}_{\Delta T})}{\Delta T} = \frac{E(\hat{X}_t) \cdot \Delta T}{\Delta T} = E(\hat{X}_t) = E(\hat{\mu}_t) \tag{6.5}$$

Where $\hat{X}_t$ represents the production volume distribution in a time unit and coincides to $\hat{\mu}_t$ since $\hat{\mu}_t$ refers to a time unit.

For what concerns the variance, equation 6.6 holds.

$$VAR(\hat{\mu}_{\Delta T}) = VAR(\frac{\hat{X}_{\Delta T}}{\Delta T}) = \frac{VAR(\hat{X}_{\Delta T})}{\Delta T^2} \tag{6.6}$$

Since it is assumed that the variance of the production is proportional to the considered ΔT period as well (i.e. $VAR(\hat{X}_{\Delta T}) = VAR(\hat{X}_t) \cdot \Delta T$ where K is a constant), equation 6.7 is obtained.

$$VAR(\hat{\mu}_{\Delta T}) = \frac{VAR(\hat{X}_t) \cdot \Delta T}{\Delta T^2} = \frac{VAR(\hat{X}_t)}{\Delta T} \tag{6.7}$$

The last equation means that the higher is the considered ΔT to estimate the production rate, the lower its variance is.

In this case, since all the computation are performed with the minutes as unit of measure, $\hat{X}_{min} = \hat{X}_t$ and $\hat{\mu}_{min} = \hat{\mu}_t$.

The estimate of the expected value of $\hat{\mu}_t$ is easy to compute because the total number of produced parts in all the 204 samples must be divided by the total required time as in equation 6.8.

$$E(\mu_{t,i}) = E(\mu_{\Delta T,t}) = \frac{\sum_{i=1}^{204} X_i}{\sum_{i=1}^{204} T_{av,100\%OEE,i}} = 2.1337 \frac{parts}{min} \tag{6.8}$$

For what concerns the variance, one estimate for each level of the non-parametric ANOVA on the number of daily shifts can be computed. In each of the three levels there can be samples which refer to slightly different ΔT but this aspect is not considered by assuming that they all refer to the same equivalent ΔT which is the mean one for that level.

Group	$\Delta T_{equivalent,i}$	DoF_i	$VAR(\hat{\mu}_{\Delta T,i})$	$VAR(\hat{\mu}_{min,i})$
1	466.7	3-1=2	0.0268649	12.5378
2	852.5	6-1=5	0.0211148	18.0004
3	1281.5	195-1=194	0.0144579	18.5278

As it can be seen, if variances of the production rate are referred to the same time period, their values become quite similar except for the first group which is estimated with few DoF. It can be now defined a pulled-estimator (see equation 6.9) to combine the three variance estimates.

$$VAR(\hat{\mu}_{min}) = \frac{VAR(\hat{\mu}_{min,1})DoF_1 + VAR(\hat{\mu}_{min,2})DoF_2 + VAR(\hat{\mu}_{min,3})DoF_3}{DoF_1 + DoF_2 + DoF_3} = 18.4545 \frac{parts^2}{min^2} \tag{6.9}$$

As a result, it is obtained:

$$E[\hat{X}_{min}] = E[\hat{X}_t] = E[\hat{\mu}_{min}] = 2.1337$$
$$VAR[\hat{X}_{min}] = VAR[\hat{X}_t] = VAR[\hat{\mu}_{min}] = 18.4545$$

The formulas have been found with adimentional values for the sake of simplicity but actually $\hat{X}_{min}$ along with its expected value are expressed in $parts$ and its variance in $part^2$. Now, since equations 6.3 and 6.4 hold, it is possible to define the weekly mean value and weekly variance of whatever production plan knowing only the weekly POT ($T_{tot,100\%OEE}$) which has been denoted as ΔT in the formulas. In particular:

$$E[\hat{X}_{plan}] = E[\hat{X}_{min}] \cdot POT \tag{6.10}$$

$$VAR[\hat{X}_{plan}] = VAR[\hat{X}_{min}] \cdot POT \tag{6.11}$$

In the presented case the POT of one shift is 430 minutes except for one shift on Tuesday, which includes only 385 minutes, and the 4 possible plans have those characteristics.

planNumber	POT $[min]$	$E[\hat{X}_{plan}]$ $[parts]$	$VAR[\hat{X}_{plan}]$ $[parts^2]$
1	5115	10914	94395
2	5975	12749	110266
3	6405	13666	118201
4	6835	14584	126137

At this point the production plans have been characterize by a weekly mean and variance and the proper Markov chain with those characteristics must be created. Since the available data about the failures only provide information about the total amount of lost time for each failure type but they do not provided information about how many times that failure occured during the day or how the MTTF is distributed and since the layout of the plant is much more complicated than a single machine, it is decided to model the exact mean and the exact variance with a Markov chain with only two states. It allows to represent an up state whose weekly rate is equal to the weekly nominal production rate. Considering figure 6.1 it can noticed that the daily OEE generally does not overcome the value of $2.3\frac{parts}{min}$ and, as a result, it is possible to set:
$\mu_{nom} = POT \cdot 2.3\frac{parts}{min}$
Another advantage on these modeling is that the upper limit for the production rate is always respected.
The MTTF and MTTR are modeled as exponentially distributed and the transition rates can be computed with the method presented in the 4^{th} chapter based on the paper [CMT10]. The obtained results are shown in the following table.

$plan$	$POT[min]$	$meanValue$	$Variance$	$maxRate$	p	r
1	5115	10914	94395	12266	34.4796	278.2495
2	5975	12749	110266	14329	40.2768	325.0324
3	6405	13666	118201	15360	43.1753	348.4239
4	6835	14584	126137	16391	46.0739	371.8153

The estimations of the rate distribution considering the real data of the third group (red line) and the equivalent results of the daily production rate computed with the simulation of the modeled Markov chain (blue line) can be compared as in figure 6.13.

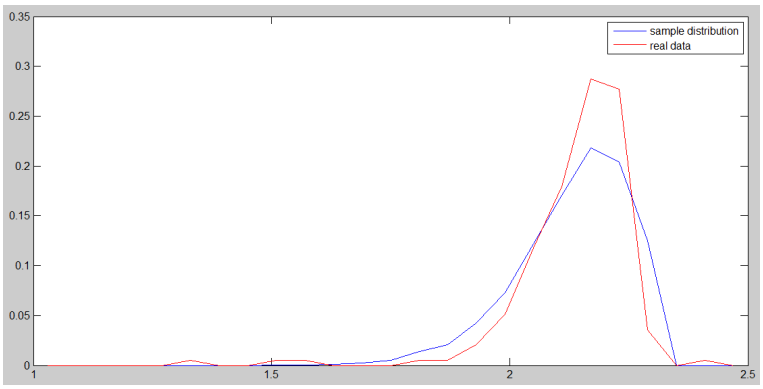

Figure 6.13: Distributions comparison

The two distributions have a similar shape but the peak of the real one seams to be higher but it must be considered that its distribution has been estimated with only 195 samples. In any case, the result is satisfactory especially because it is obtained with only two states.

6.1.4 Decision delay modeling

The decision delay must model the mean reaction delay when a threshold is crossed. Although the decision is taken every two weeks, it is also possible to schedule different plans for the first and the second week. The first plan is decided knowing the exact inventory level at the beginning of the first week but its value at the beginning of the second week is known only with uncertainty, which derives from both the production variability and the demand variability. However, it is experienced that this estimate is generally quite accurate for the considered production context if no disruptive events occur.

As a result, it can be modeled that at the beginning on each week the system can be reconfigured differently, which implies a different production plan. Since threshold crossings can occur anytime during the week and the system must wait the beginning of the next week to be reconfigured, the reconfiguration delay is between 0, if the threshold crossing occurs at the end of the week, and 1 week, if the threshold crossing occurs in the first shift on Monday. It means that the delay is uniformly distributed with minumum value equal to 0 and maximum value equal to 1 week, which correspond to an expected value of half a week.

Since our model is only able to deal with exponentially distributed delays, this distribution is used instead of the uniform one and it implies an approximation. In order to model at least the correct expected value, λ is chosen equal to $\frac{1}{0.5week}$.

6.2 Escalation levels

Currently the escalation levels are defined with respect to the Kanban number NK and, precisely, as percentage of the total number of products which can be associated to all the cards. This maximal quantity is obtained by multiplying the number of Kanbans by the mean number of parts associated to one card NPK. Since the number of Kanbans is changed every 2 weeks the exact maximal inventory level can only be estimated. As already discussed the NK already depends on all the possible variabilities of the production plant and of the demand and, precisely, the more variable the system is, the bigger their number is because the stockout risk increases.

The maximal inventory level is not set to 100% of the Kanban number but to 80% because in practice, when the 80% of Kanbans cards are already in the warehouse, the plant is already unable to produce by their absence at the beginning of the line.

However, the company assumes that the number of parts per Kanban (NPK) is equal to 50 on average and the total number of Kanban is equal to 600 and, as a consequence, the product $NPK \cdot NK$ is equal to 30000 on average. The escalation levels are computed considering this product. In particular, the 16-shifts plan is used if the inventory level is lower than $30\% NPK \cdot NK$, the 15-shifts plan if it is between 30% and $50\% NPK \cdot NK$, the 14-shifts plan if it is between 50% and $60\% NPK \cdot NK$ and, finally, the 12-shifts plan if it is greater than $60\% NPK \cdot NK$ (figure 6.14).

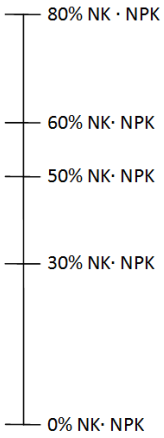

Figure 6.14: Escalation level

If it is assumed $NPK \cdot NK$=30000 the thresholds in figure 6.15 are obtained.

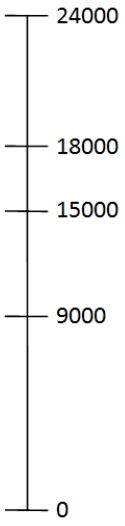

Figure 6.15: Escalation level

However, the total available buffer capacity which could be exploited is 30000 *parts* and, for this reason, this value will be considered for the maxInv parameter during the optimization (see chapter 4).

6.3 Downstream machine modeling

With the methodology described in chapter 4 it is possible to model different demand scenarios and the transitions between them. In this real case, it is possible to relate

each daily scheduled demand quantity along with its variability to each scenario and to model that on average every 24 hours the demand scenario can change. In this section, since no demand data have been provided by the company, the general method to model the downstream machine starting from a forecast demand is developed and used for a typical demand profile created with the available information.

6.3.1 Considered demand profile

The available data during the biweekly meeting are the daily demands of the next month but, since the decision only concerns a two weeks period and after that a new decision can be taken, only the data of the first and the second week are taken into consideration to model the customer behavior.

Since no data have been provided for this case, a typical demand profile is created by considering the usual known demand characteristics. It is known that the mean daily demand is approximately 2600 $\frac{parts}{day}$ and its real values can vary between $\pm 40\%$ due to the fact that most of the customers plan to take products regularly but not every day.

Furthermore, each daily quantity is not known precisely because the customer are not forced to strictly retrieve parts as scheduled in advance. It can be generally assumed that the expected value of the required quantity is equal to the scheduled demand it is normally distributed with standard deviation equal to $\sigma = \frac{\alpha \cdot scheduledDemand}{1.96}$. α is assumed equal to 0.2 because the company assumes that with a probability 95% the real demand will in the range $[(1-\alpha) \cdot scheduledDemand, (1+\alpha) \cdot scheduledDemand]$.

A typical scheduled demand series for a two weeks period is considered. The daily expected values are in the range $(1560, 3640)$, which corresponds to 2600 $\frac{parts}{day}$ $\pm 40\%$. After that, the error bars in figure 6.16 represent the range $[expectedValue - 1.96\sigma, expectedValue + 1.96\sigma]$.

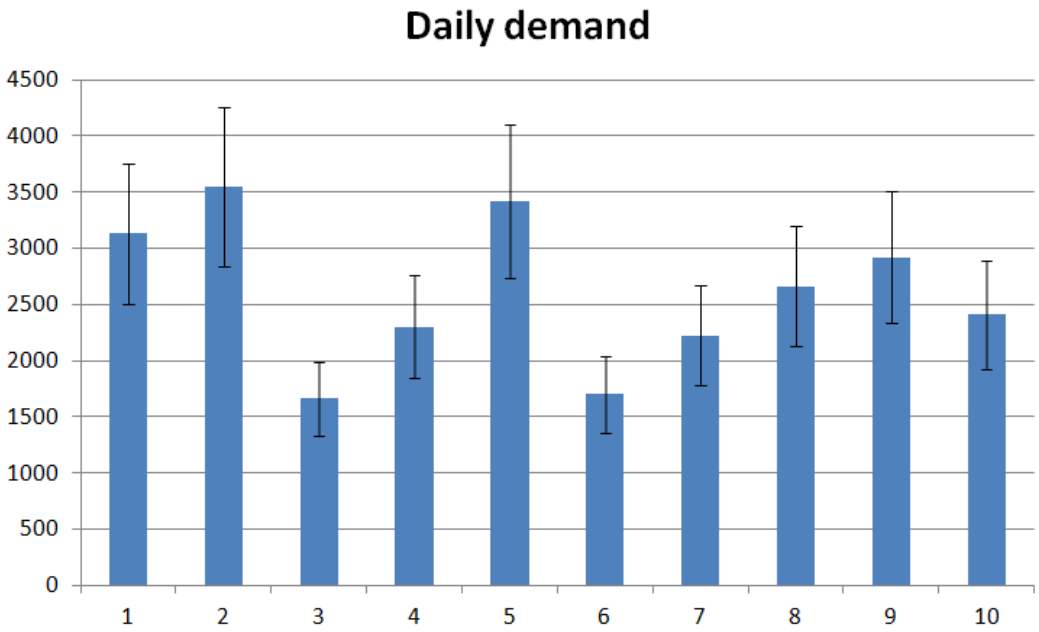

Figure 6.16: Reference demand scenario

	$Monday1$	$Tuesday1$	$Wednesday1$	$Thursday1$	$Friday1$
$1^{st}week$	3130	3549	1660	2302	3421

	$Monday2$	$Tuesday2$	$Wednesday2$	$Thursday2$	$Friday2$
$2^{nd}week$	1703	2225	2664	2921	2408

6.3.2 Modeling of the scenarios

It is possible to group the similar demands to obtain a limited number of scenarios as in figure 6.17 and to maintain fast computation times. In particular, it is considered that the similar demand quantities belong to the same scenario.

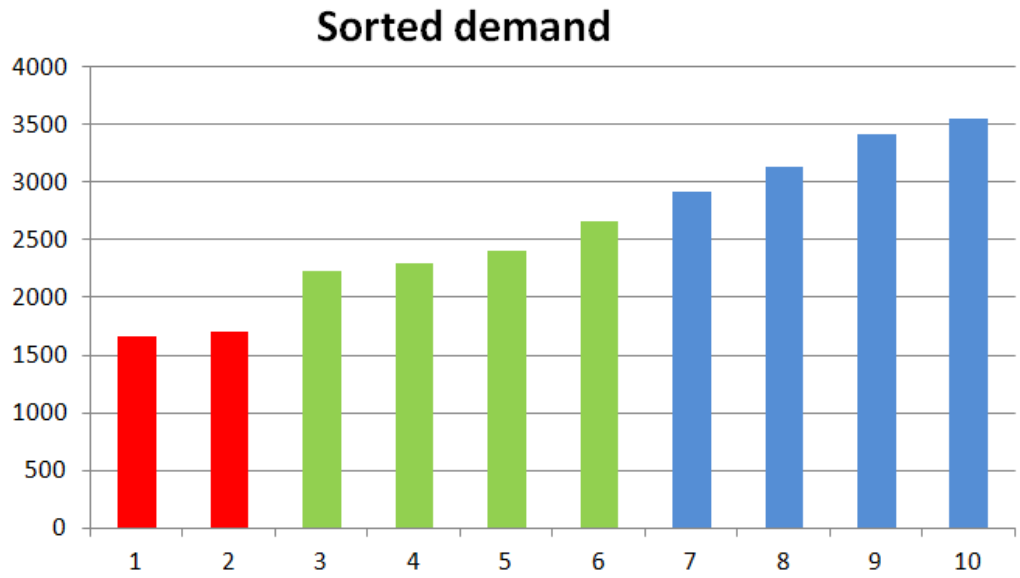

Figure 6.17: Sorted demand

	$Wednesday1$	$Monday2$
$1^{st} scenario$	1660	1703

	$Thursday1$	$Tuesday2$	$Wednesday2$	$Friday2$
$2^{nd} scenario$	2302	2225	2664	2408

	$Monday1$	$Tuesday1$	$Friday1$	$Thursday2$
$3^{rd} scenario$	3130	3549	3421	2921

Three groups of demand are defined and each of them is refers to a scenario demand. Since the small computation time, which is strongly dependent on the number of used states, is a priority for our model, it is chosen to represent each scenario with a Markov chain (see figure 6.18) that includes only two states (1 up-state with production rate μ_{nom} and 1 down-state).

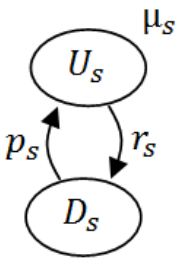

Figure 6.18: Markov chain of the s^{th} demand scenario

The parameters of the Markov chain of each s^{th} scenario (μ_s, p_s and r_s) can be computed as follows:

1. The scenario expected value of the demand is computed as the mean of the daily demands of all the days which belong to that group and it is assumed as the scheduled demand for that scenario ($scheduledDemand_{scenario}$).

2. Since it is assumed that the real demand will not be greater than $1.2 \cdot scheduledDemand_{scenario}$ ($\alpha = 0.2$) with probability 97.5%, it is assumed that $\mu_{nom,scenario}$, which represents the maximum possible production rate, is equal to $1.2 \cdot scheduledDemand_{scenario}$. the rate of the up-state of the scenario μ_{nom} is set as 1.2 chosen as the maximum value of the range where there is the 95% of the probability and in this case 1.2 times the weekly expected value since it is assumed that the planned quantity of each day is known with an uncertainty of ±20%

3. The scenario probability represents the frequency in the 2 weeks period of one demand group divided by the total number of considered days.
$P_{scenario} = \frac{numberOfDays_{scenario}}{totalNumberOfDays}$

4. The variance on a daily basis can be computed as follows:

$$dailyVariance = \frac{(\alpha \cdot dailyExpectedValue)^2}{1.96^2} \tag{6.12}$$

5. The weekly scenario variance, expected value and production rate of the up-state can be computed multiplying the daily values by 5, i.e. the number of days in a week.

6. Once those three weekly values are known, it is possible to use the method based on [CMT10] to compute the correspondent p and r production rates.

The obtained values for the given demand profile are:

	$group1$	$group2$	$group2$
$dailyexpectedvalue$	1682	2400	3255
$weeklyexpectedvalue$	8408	11999	16276
$P_{scenario}$	0.2	0.4	0.4
μ_{nom}	10089	14399	19532
$Weeklyvariance$	147201	299813	551679
p	32.01	32.01	32.01
r	160.07	160.07	160.07

At this point the sub-Markov chains of all scenarios are defined and it is noticed that, since the production rates and the standard deviation of each scenario are proportional to the expected value, the p and r rates are equal in all scenarios.

6.3.3 Modeling of the scenario transitions

Furthermore, it must be defined how and how often it is possible to switch from one scenario to the others, which are characterized by the same p_s and r_s but different μ_s. It can be assumed that the demand scenarios are independent, namely that on average every day each demand scenario can occur with its own probability $P_{scenario}$ and it is independent from the previous one. Since the transition rates are big and in order to maintain a simple Markov chain layout, it is decided to model the transitions as if the up states and the down states of the downstream machine represent real states, i.e. the transitions between scenarios are not randomized. The resulting Markov chain is presented in the figure 6.19.

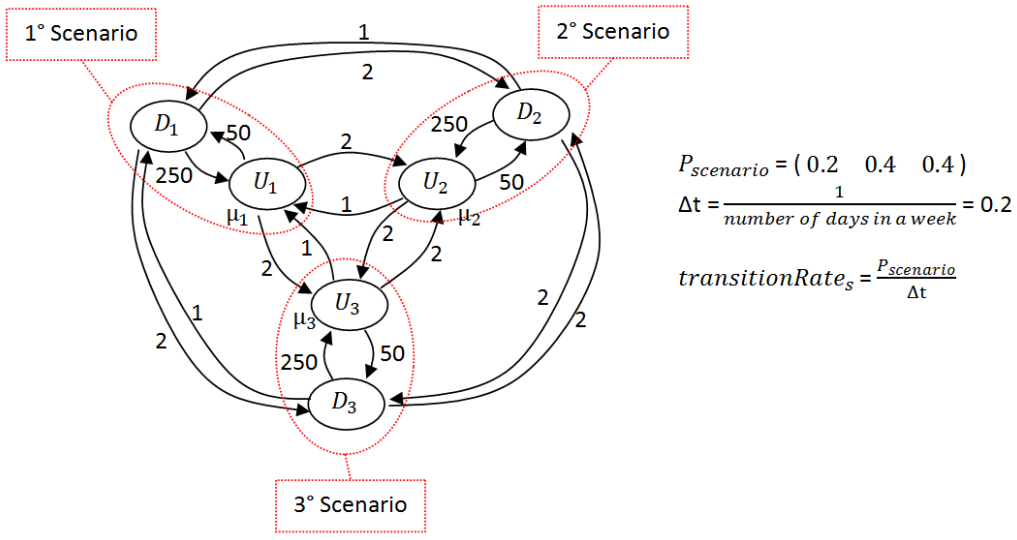

Figure 6.19: Markov chain downstream machine

6.4 Cost evaluation and optimization

Once the modeling of the real system is done, the computation can be run and the outputs which are required for the cost evaluation, i.e the mean inventory level, the mean backlogged quantity and the probability associated to each production configuration, are obtained and they can be used in the objective function (6.13).

$$OF = avInv \cdot C_{finishedProduct} \cdot r_{stock} + Q_{backlog} \cdot C_{backlog} + P_1 \cdot C_1 + P_2 \cdot C_2 + P_3 \cdot C_3 + P_4 \cdot C_4 \tag{6.13}$$

6.4.1 Cost coefficient

In order to use the equation 6.13 is necessary to determine the cost coefficient which convert a steady-state index in a cost related to a defined period, which in this case it is the week.

- $C_{finishedProduct}$ represents the production cost of a product which is stocked in the final warehouse, its measurement unit is $\frac{euro}{part}$ and, when it is multiplied by $avInv$, the product represents the mean value of the total inventory. In this case, each final product has an industrial cost of $110\frac{euro}{part}$

- r_{stock} is expressed in $\frac{1}{week}$ and represents the weekly stocking cost for an inventory value of 1 $euro$. Since the yearly r for the company is assumed to be 0.08 and there 50 working weeks in year, the weekly $r_{stock} = 0.0016$

- $C_{backlog}$ is expressed in $\frac{\frac{euro}{week}}{part}$ and it represents the cost that the company incurs when 1 product remains in backlog for 1 week. This coefficient is difficult to be determined precisely but a reasonable value, which takes into account not only the costs fines defined in the contract but also the image loss and the fact that the customer can change the supplier if it occur often, is $C_{backlog} = 500\frac{\frac{euro}{week}}{part}$

- The considered weekly cost of each production configuration concerns only the manpower costs. There are of course other operative costs, for instance material costs, energy costs, etc.. but they are assumed proportional to the produced quantities and, since the demand is never lost, they are not differential.
  Let us start computing the cost of the normal plan which is the one with 15 shifts. In each shift there are 120 workers in the department plus 8 shift supervisors which has the same wage. The manpower cost for the company is $40\frac{euro}{hour \cdot worker}$. This cost must be increased by 50% if the workers are employed between 6 p.m. and 6 a.m. or on Saturday. The first shift is between 6 a.m. and 2 p.m. and, as a result its cost is $40\frac{euro}{hour \cdot worker}$. The second one is between 2 p.m. and 10 p.m. and its costs will be $50\frac{euro}{hour \cdot worker}$ because it is half in the ordinary time and half in the evening. Finally, the third one is between 10 p.m. and 6 a.m., i.e. completely in the night,

and its cost is $60\frac{euro}{hour \cdot worker}$.
As a result, the cost of the normal plan is defined in equation 6.14.

$$C_3 = 15\frac{shifts}{week} \cdot 8\frac{hours}{shift} \cdot 128 workers \cdot 50\frac{euro}{hour \cdot worker} = 768000\frac{euro}{week} \quad (6.14)$$

In this case the manpower cost is $50\frac{euro}{hour \cdot worker}$ because the 15-shifts plan includes 5 shifts in the morning, which cost $40\frac{euro}{hour \cdot worker}$, 5 shifts in the afternoon, which cost $50\frac{euro}{hour \cdot worker}$ and 5 shifts in the night, which cost $60\frac{euro}{hour \cdot worker}$.

Now the C_1, C_2 and C_4 must be computed.
The plan with 16 shifts is obtained by adding a shift on Saturday and it implies the costs of 8 hours in overtime.
$\Delta costs_{C_4} = +1 shift \cdot 8\frac{hours}{shift} \cdot 128 workers \cdot 60\frac{euro}{hour \cdot worker} = +61440\frac{euro}{week}$
On the contrary, when some shifts must be removed, it is considered that there are 3 possible solutions.

1. It is possible to ask some workers to request vacation days, if the still have. This solution involves no costs for the company.
2. The costs can be absorbed by the social welfare, namely that the state pays for it, and the company incurs no costs.
3. The company pays the salaries but it makes them work on other departments. In this case the manpower will not be always used efficiently because the tasks can be different and the law of diminishing returns holds. For example, if another department is designed to work with the current amount of manpower and it is increased by 10%, the throughput generally also increases but by a lower percentage.

As a consequence, only a portion of their costs can be assigned to the other lines. Furthermore, if only one shift is removed, it is possible to delete the last shift on Friday which is the one which costs more. On the contrary, if 3 shifts must be removed and since they are all removed on Friday to maintain the production continuous, also the some other shifts are removed

In any case, it is assumed that, in both cases, the three solutions occur which those probabilities:

1. 20% the workers request vacation days
2. 50% the aids of the social welfare are used
3. 30% the workers are paid and employed on other lines with efficiency of 60%

It means that on average for C_2:
$\Delta costs_{C_2} = -(0.2 + 0.5 + 0.3 \cdot 0.6) \cdot 1 shift \cdot 8\frac{hours}{shift} \cdot 128 workers \cdot 60\frac{euro}{hour \cdot worker} = -54067\frac{euro}{week}$

and for C_1:
$\Delta costs_{C_1} = -(0.2 + 0.5 + 0.3 \cdot 0.6) \cdot 3shift \cdot 8\frac{hours}{shift} \cdot 128workers \cdot 50\frac{euro}{hour \cdot worker} = -135168\frac{euro}{week}$

To sum up, the following weekly manpower costs are incurred with the 4 possible plans.

$planNumber$	$\Delta Cost$	C_i	$producedquantities$	$unitaryCost$
1	-135168	632832	10914	57.98
2	-54067	713933	12749	56.00
3	0	768000	13666	56.20
4	+61440	829440	14584	56.87

If the unitary costs are diagrammed (figure 6.20), it is possible to notice that the most cost-efficient policies are the second and third one. Since the mean weekly demand is $13000\frac{part}{week}$, if the demand is not very turbulent, the system switches between those two plans. On the contrary, if the demand becomes very turbulent, also the two extreme policies are used to avoid great backlog and inventory costs.

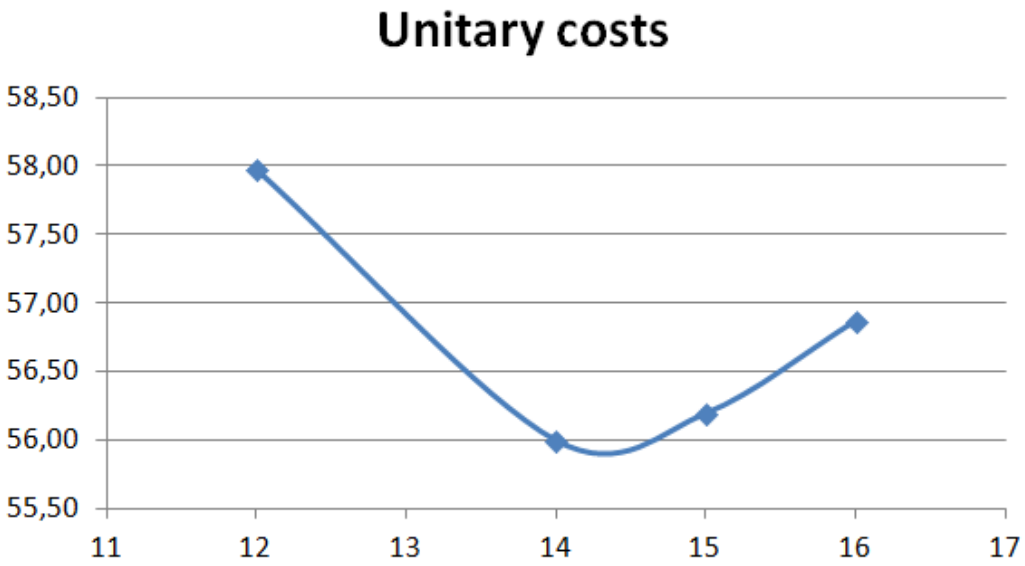

Figure 6.20: Unitary costs of the plans

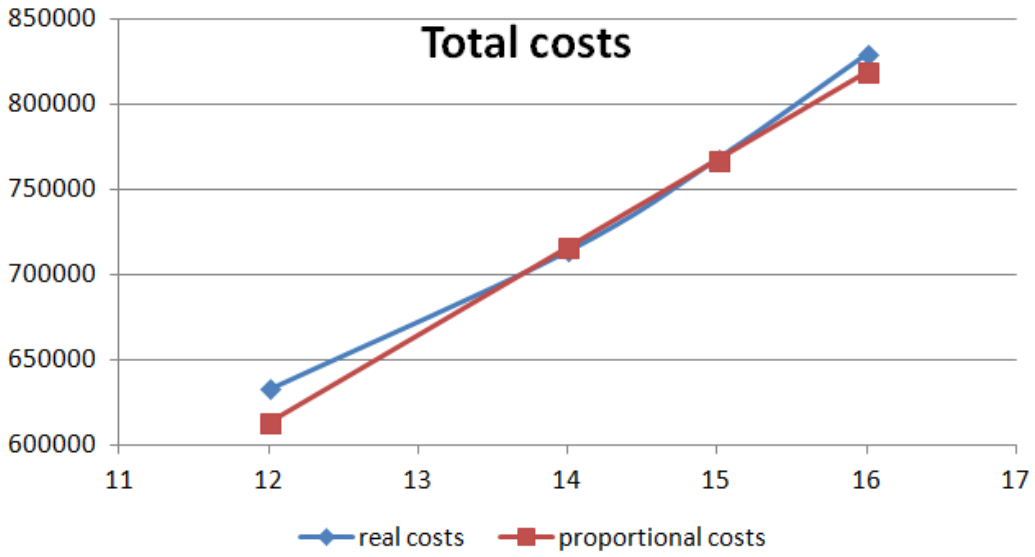

Figure 6.21: Total costs of the plans

In conclusion, the objective function will be as follows:

$$OF = avInv \cdot 110\frac{euro}{part} \cdot 0.00625\frac{\frac{euro}{week}}{euro} + Q_{backlog} \cdot 500\frac{euro \cdot part}{week} + P_1 \cdot 632832\frac{euro}{week} + \\ + P_2 \cdot 713933\frac{euro}{week} + P_3 \cdot 768000\frac{euro}{week} + P_4 \cdot 829440\frac{euro}{week}$$

6.4.2 Optimization

It is now possible to launch the cost evaluation and the obtained results are shown in the following table.

$avInv$	14835 $parts$
$Q_{backlog}$	0.024 $parts$
P_1	0.130
P_2	0.380
P_3	0.455
P_4	0.035
$P_{lowerThreshold}$	$\simeq 0$
$C_{manpower}$	$732065\frac{euro}{week}$
$C_{backlog}$	$12\frac{euro}{week}$
$C_{inventoryCosts}$	$2611\frac{euro}{week}$
C_{tot}	$734688\frac{euro}{week}$
$t_{computation}$	$25sec$

At this point the optimization can be run to discover if there is a better reaction policy for this precise plant and demand behavior.
Basically, the optimization problem has only 4 unknows, i.e. x_1,x_2,x_3 and x_4, because the width of the negative buffer is kept constant and 5 constraints which are:

$$\begin{cases} N_1 > N_0 \\ N_2 > N_1 \\ N_3 > N_2 \\ N_4 > N_3 \\ N_4 \leq invMax \end{cases} \tag{6.15}$$

The first four constraints assure that the order of the escalation threshold remains the same and that the width of each range can not be zero. However, if the optimal thresholds include some layers which have a width which is approximately zero, it is probably better to delete those ranges and re-run the optimization. The $invMax$ represents the physical maximal capacity of the buffer which can not be exceeded. All the three possible algorithms are used and all return the same solution but different computation times.

$algorithm$	$T_{computation}$
$Interiorpoint$	$5536sec$
SQP	$5273sec$
$Active-set$	$4850sec$

Since the active-set algorithm seems to be the less time consuming for this optimization, it will be used from now on.
The optimization algorithm implemented in Matlab returns the following optimal reaction thresholds:

N_i	$thresholdLevel_{currentPolicy}$	$thresholdLevel_{optimizedPolicy}$
N_4	24000	30000
N_3	18000	26414
N_2	15000	9664
N_1	9000	5275

It can be noticed that the 14-shifts plan has a larger range in the optimal reaction policy. It happens mainly for two reasons. Firstly, it (along with the 15-shifts plan) is the more cost-efficient and, secondly, its expected value is quite similar to the expected value of the demand.

Secondly, it is better to use the entire available buffer capacity, instead of limiting it with the number of Kanbans, because the OEE losses due to the lack of production cards will be lower.

Important considerations can be done by considering also other model outputs, shown in the following table:

	$currentPolicy$	$optimizedPolicy$
$avInv$	14835 $parts$	13539 $parts$
$Q_{backlog}$	0.024 $parts$	0.47 $parts$
P_1	0.130	0.014
P_2	0.380	0.724
P_3	0.455	0.228
P_4	0.035	0.033
$P_{lowerThreshold}$	$\simeq 0$	$\simeq 0$
$C_{manpower}$	$732065\frac{euro}{week}$	$728929\frac{euro}{week}$
$C_{backlog}$	$12\frac{euro}{week}$	$234\frac{euro}{week}$
$C_{inventoryCosts}$	$2611\frac{euro}{week}$	$2383\frac{euro}{week}$
C_{tot}	$734688\frac{euro}{week}$	$731546\frac{euro}{week}$

The computation of the optimal policy underlines that the current policy is too reactive because it focuses on limiting the backlog costs as much as possible by making the manpower costs increase. Indeed, it is better to make the buffer absorb more the demand fluctuation and to react only if the buffer becomes extremely low (backlog risk) or extremely high (risk that the plant is forced not to product due to the blocking caused by the full buffer or by the lack of Kanban cards).

The new value for the objective function is $731545\frac{euro}{week}$ and, if it is compared to the current value of $734688\frac{euro}{week}$, it can be concluded that, according to the model, it is possible to save $3143\frac{euro}{week}$, which corresponds to $157150\frac{euro}{year}$ if the company uses the optimized policy instead of the current one.

6.5 Biweekly decision delay

The model developed in this chapter behaves as if the reconfiguration decision could be taken every week. However, it is possible to take plan decision only every two weeks and, in particular, when the biweekly meeting is organized, both the production plan for the incipient week (called first week) and for the following one (called second week) are defined and they can not be changed until the next biweekly meeting. It implies that the decision delay can vary between 0, if the threshold crossing occurs and the end of the second weekly, i.e. immediately before the next meeting, and 2 weeks, if the threshold crossing occurs at the beginning of the 2-weeks period, i.e. immediately after the meeting.

If we would like to model this situation, it is possible to set a delay whose expected value is equal to 1 week (the double as in the previously modeling) and to define that during the meeting the total number of shifts for the 2-weeks period is chosen. On a weekly basis, the system can use only 4 production plans (12,14,15,16 shifts) but, if two weeks are considered together, all the possible combinations must be considered.

$totalShiftNumber$	$availableCombinations(shifts_{1^{st}week} + shifts_{2^{nd}week})$
$24shifts$	$(12+12)$
$26shifts$	$(12+14),(14+12)$
$27shifts$	$(12+15),(15+12)$
$28shifts$	$(12+16),(14+14),(16+12)$
$29shifts$	$(14+15),(15+14)$
$30shifts$	$(14+16),(15+15),(16+14)$
$31shifts$	$(15+16),(16+15)$
$32shifts$	$(16+16)$

8 possible configurations are identified (24,26,27,28,29,30,31,32 $\frac{shifts}{2weeks}$) but this modeling is not able to understand the differences between two configurations with the same number of shifts (for example, 12 shifts on the first week and 14 on the second one denoted as (12+14) or viceversa (14+12)).

However, a model with 8 configurations a larger number of states on each layer, 8 ranges + 1 negative range and 8 threshold to optimize. The optimization of such a system takes in practice a too large amount of time and the computation problems becomes much more likely, mainly due to the overflow of some variables. Indeed, the Matlab code uses a multiprecision toolbox [Mul14] which allows the code to use the quadruple precision variables to deal with a larger exponent range (10^{-4965} ; 10^{4932}). This exponent range is enough to deal with modeling based on the weekly decision but it is generally not enough to deal with this 9-threshold model.

That is the main reason why all the modeling and the computation has been done with the simpler 4-thresholds model although it considers a different mean decision delay.

6.5.1 Delay length comparison

Although it is not possible to model a biweekly decision with those 4 possible plans because it implies 8 thresholds to optimize, it is possible to model a situation where the decision is taken every two weeks but the available weekly configuration are only 12, 14 and 16 shifts (without the 15-shifts policy).

Also in this case, the managers must take two decisions during the biweekly meeting but with only three alternatives every week. The resulting combinations are:

$totalShiftNumber$	$availableCombinations(shifts_{1^{st}week} + shifts_{2^{nd}week})$
$24shifts$	$(12+12)$
$26shifts$	$(12+14),(14+12)$
$28shifts$	$(12+16),(14+14),(16+12)$
$30shifts$	$(14+16),(16+14)$
$32shifts$	$(16+16)$

Only 5 possible configurations are identified (24,26,28,30,32 $\frac{shifts}{2weeks}$) but this modeling is also not able to appreciate the differences between two configurations with the same number of shifts (for example, 12 shifts on the first week and 14 on the second one denoted as (12+14) or viceversa (14+12)).

The Matlab code can easily handle the modeling of 5 possible configurations, 5 ranges + 1 negative range and 5 thresholds to optimize.

This modeling make us able to estimate the economical advantage of planing the number of shifts weekly and not every two weeks when only those three production plans are available (12,14 and 16 shifts). In particular, two models must be compared. They include the same customer behavior but the upstream machine and number of thresholds are different.

Model 1 - weekly meeting The first one models that every week the managers decide the number of shifts to plan for that week and they have three possibilities: 12, 14 and 16 shifts. As a consequence, 3 different configurations and 3 ranges plus a negative one must be modeled. The number of ranges to optimize is 3 as well and, since the plan decision is taken every week, the expected value of the reaction delay when a threshold is crossed is half a week.

$shiftNumber\ [\frac{shifts}{week}]$	$POT\ [\frac{min}{week}]$	$E[\hat{X}_{week}]\ [parts]$	$VAR[\hat{X}_{week}]\ [parts^2]$	$costs\ \frac{euro}{week}$
$12shifts$	5115	10914	94395	632832
$14shifts$	5975	12749	110266	713933
$16shifts$	6835	14584	126137	829440

The values of POT, $E[\hat{X}_{week}]$, $VAR[\hat{X}_{week}]$ and costs for the considered policies have already been computed in previous sections of this chapter and, therefore, they do not have to be computed again.

Model 2 - biweekly meeting The second one models that every 2 weeks the managers decide the total number of shifts to plan for both weeks and they have five possibilities: 24, 26, 28, 30 and 32 shifts. As a consequence, 5 different configurations and 5 ranges plus a negative one must be modeled. The number of ranges to optimize is 5 as well and, since the plan decision is taken every 2 weeks, the expected value of the reaction delay when a threshold is crossed is 1 week.

$shiftNumber\ [\frac{shifts}{2weeks}]$	$POT\ [\frac{min}{week}]$	$E[\hat{X}_{week}]\ [parts]$	$VAR[\hat{X}_{week}]\ [parts^2]$	$costs\ \frac{euro}{week}$
$24shifts$	5115	10914	94395	632832
$26shifts$	5545	11831	102330	673382
$28shifts$	5975	12749	110266	713933
$30shifts$	6405	13666	118201	771686
$32shifts$	6835	14584	126137	829440

Also in this case, the values of POT, $E[\hat{X}_{week}]$ and $VAR[\hat{X}_{week}]$ for the considered policies have already been computed in previous sections of this chapter. In particular, the plan with $\frac{24shifts}{2week}$ has the same values of the 12-shifts plan, the plan with $\frac{26shifts}{2week}$ has the same values of the 13-shifts plan, the plan with $\frac{28shifts}{2week}$ has the same values of the 14-shifts plan and so on. Therefore, they do not have to be computed again. The only values which must be computed are the weekly costs for each plan:

$$costs_{plan} = \frac{costs_{plan\ 1^{st}week} + costs_{plan\ 2^{nd}week}}{2}\ [\frac{euro}{week}]$$
$$costs_{24shifts} = \frac{costs_{12shifts} + costs_{12shifts}}{2}\ [\frac{euro}{week}].$$
$$costs_{26shifts} = \frac{costs_{12shifts} + costs_{14shifts}}{2}\ [\frac{euro}{week}].$$
$$costs_{28shifts} = \frac{costs_{14shifts} + costs_{14shifts}}{2}\ [\frac{euro}{week}].$$
$$costs_{30shifts} = \frac{costs_{14shifts} + costs_{16shifts}}{2}\ [\frac{euro}{week}].$$
$$costs_{32shifts} = \frac{costs_{16shifts} + costs_{16shifts}}{2}\ [\frac{euro}{week}].$$

The obtained optimal thresholds are shown in figure 6.22.

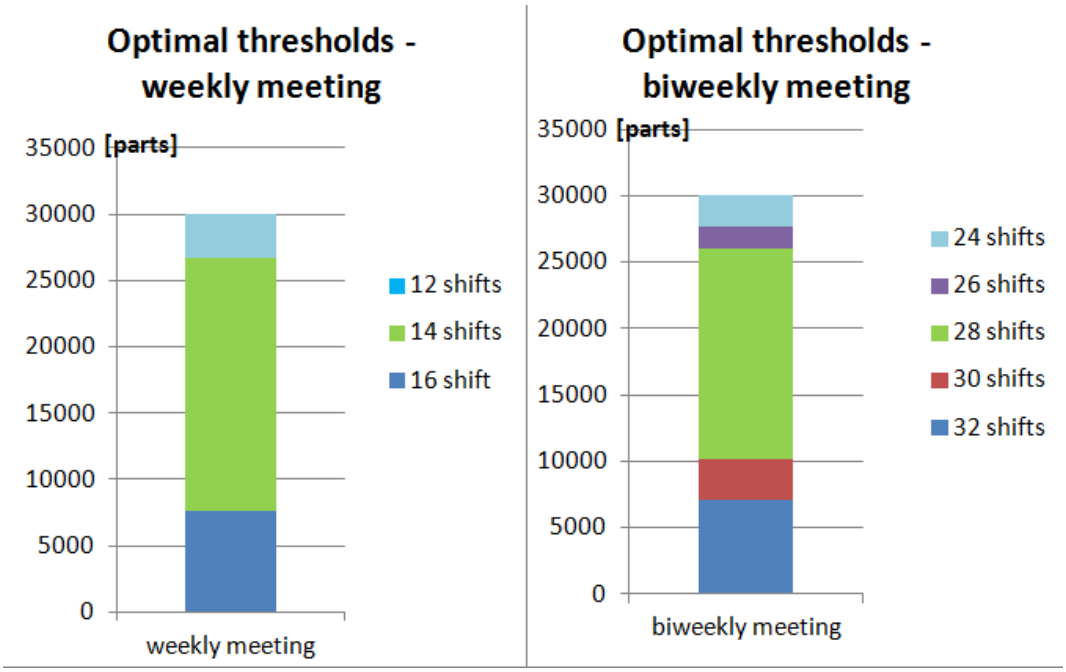

Figure 6.22: Optimal thresholds for both models

It can be seen that, if the production plan is decided every week, the optimal reaction policy becomes less reactive because the 14-shifts plan, which is the most cost-efficient, is used for a wider buffer range.

	$Optimized\ OF[euro/week]$
$Weeklymeeting$	732273
$Biweeklymeeting$	732776

It is estimated that, if the production plans are decided every week instead of every 2 weeks, a saving of $503\frac{euro}{week}$, which corresponds to $25150\frac{euro}{year}$ can be achieved. For this reason, it is strongly suggested to plan the number of shifts every week and to use the

model defined in the first part of this chapter, which considers 4 configurations (12,14,15 and 16 shifts) and has a mean decision delay of half a week.

7 Optimal thresholds considerations

Since the optimal ranges must be recomputed every time a new meeting takes place and generally the production and demand characteristics can vary during the whole year, also the optimal escalation levels will be different. As a result, in this chapter it is shown how they vary depending on the production context and how the costs are badly affected if the thresholds are kept fixed. Moreover, since the company produces in different countries where the inventory, manpower and backlog costs, the power of the union, the manpower flexibility and so on are different, it is also shown which effects they have on the optimal escalation levels. Finally, it is shown that, in this real case, since the proper reaction policies have been wisely chosen and assigned to the proper ranges, it is generally better to have more escalation levels which implies stronger reactions when the inventory level is far from the normal range. In all these experiments the objective function is also evaluated to understand how those aspects affect it.

In each experiment the situation modeled and studied in the first sections of the previous chapter, which is assumed and denoted as "reference case", is always taken into consideration. In particular, the computation and the results are obtained using the model with 4 available plans (12,14,15 and 16 shifts) and an average reaction delay of half a week, namely that the plan can be decided every week by evaluating the current buffer. As already discussed, in the previous chapter, it is not exactly the current decision policy used in Bosch but it is the one that is suggested to the company (the economical advantage of changing the decision-making method has been evaluated in the previous chapter).

Starting from the reference case it is shown what happens if a single characteristic or objective function parameter at a time changes. Moreover, it is quantified how much money the company would lose if the escalation thresholds were not changed but kept fixed and, precisely, equal to the optimal ones for the reference case. Those considerations have mainly two purposes: quantify the advantage of adapting the reaction policy to the current production environment and, secondly, to understand the importance of estimating the input parameters for the model correctly. Indeed, if the current production environment (plant and customer) was not modeled correctly, the computed reaction thresholds would not be the optimal ones.

7.1 Cost coefficients

In this first section the coefficients of the objective function, which transforms the performances of the system in costs, are varied. Let us briefly recall the objective function previously defined in 6.13.

$$OF = avInv \cdot C_{finishedProduct} \cdot r_{stock} + Q_{backlog} \cdot C_{backlog} + P_1 \cdot C_1 + P_2 \cdot C_2 + P_3 \cdot C_3 + P_4 \cdot C_4 \quad (7.1)$$

7.1.1 Inventory costs

The first considered cost parameter is the product $C_{finishedProduct} \cdot r_{stock}$ and it can vary for two reasons:

1. if the $C_{finishedProduct}$ varies. It can happen if the product mix changes (it has been considered an equivalent product but in general all the products have different production costs) or if new products are produced

2. if the r_{stock} varies. As already said, it considers two aspects, i.e. the stocking of goods costs money because the warehouse must be built, managed, the goods can become obsolete, etc.. and the value of the stocked goods can't be reinvested. As a result, if a new investment opportunity for the company with a high internal rate of return (IRR) occurs, the r_{stock} must be increased because the value of the stocked goods could be invested in a more profitable way and it is better to limit the average inventory level even more.

Figure 7.1 shows how the optimal reaction levels and the corresponding objective function change with respect to that coefficient. In particular, on the horizontal axis there is the ratio (expressed as a percentage) between the inventory coefficient and the one of the real case which is equal to $0.176 \frac{euro}{part \cdot week}$.

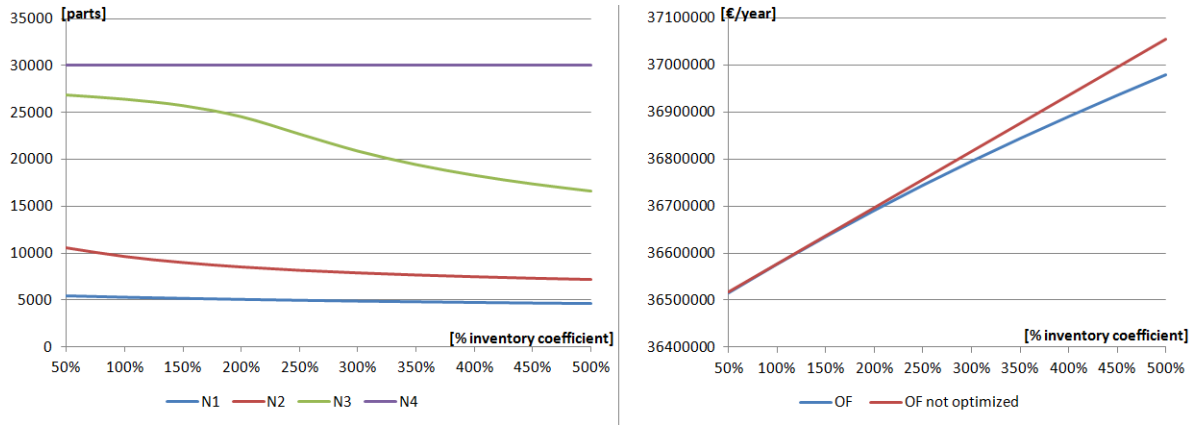

Figure 7.1: Optimal thresholds and OF as a function of the inventory cost

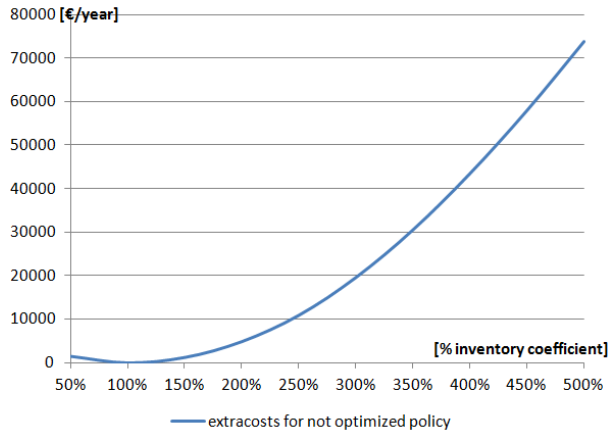

Figure 7.2: Extra costs if the policy is not changed

Since all the cost coefficients remain the same except for the considered one which increases, the optimal value of the objective function also increases. The greatest effect is on the third threshold which decreases faster than the others because, in order to decrease the average inventory level, the escalation policy which removes shifts must react before the inventory level becomes too big. The upper threshold remains at the maximum level because in the considered range it is always preferable to incur some little extra inventory costs than blocking the production before the inventory is full. However, at a certain point, if the stocking costs become even bigger, the upper threshold will also decrease but it is unlikely that they become more than five times the current value, in a practical situation.

It is also interesting to show how much extra cost the company incurs if the optimal policy for the reference case is not changed when the inventory coefficient varies. As it can be already seen in figure 7.1 and more precisely in figure 7.2, the more the inventory coefficient is different from the one of the reference case, the bigger the extra costs are. That happens basically because the reference case optimal policy is used for a scenario more and more different.

7.1.2 Backlog costs

The second considered cost parameter is the $C_{backlog}$ which in the real case is $500\frac{euro}{part \cdot week}$ but it must be changed if, for instance, the customers define higher or lower backlog fine or if the company think that the image loss of the company associated to a backlog situation can be higher than before.

Figure 7.3 shows how the optimal reaction levels and the corresponding objective function change with respect to that coefficient. In particular, on the horizontal axis there is the ratio between the cost coefficient and the one of the real case which is equal to $500\frac{euro}{part \cdot week}$.

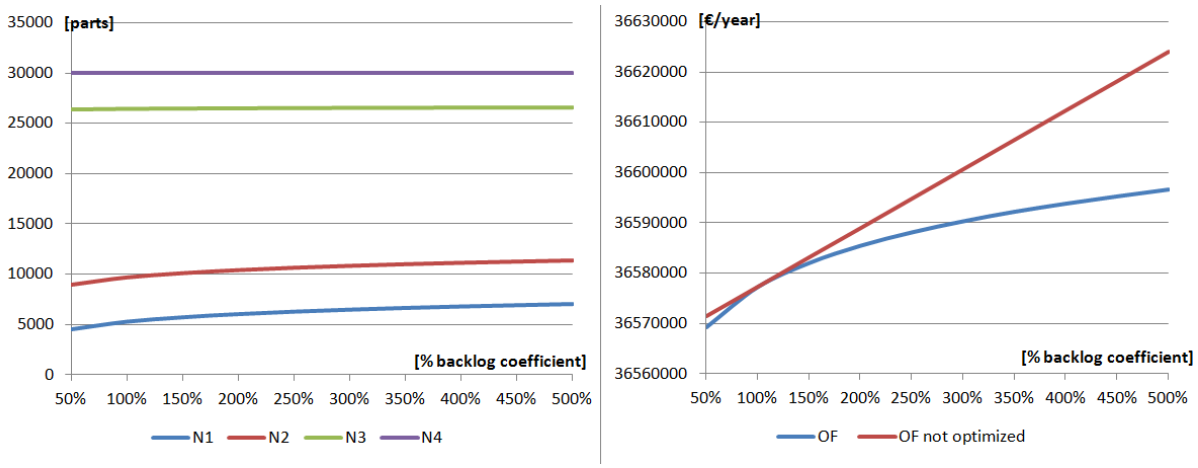

Figure 7.3: Optimal thresholds and OF as a function of the backlog cost

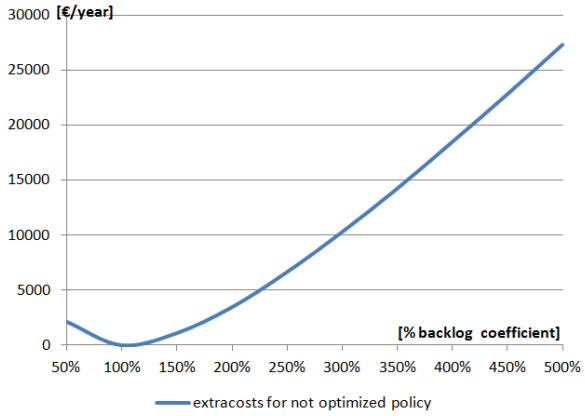

Figure 7.4: Extra costs if the policy is not changed

Since all the cost coefficient remain the same except for the considered one which increases, the optimal value of the objective function also increases. The great effects are noticed on the first and the second threshold which increase because they are the ones which make the production volumes increase as the inventory level gets dangerously low and the system is made react before to limit the backlog risk.

As for the inventory costs, it is interesting to show how much extra cost the company incurs if the optimal policy for the reference case is not changed when the backlog coefficient varies and used in other scenarios. As it can be already seen in figure 7.3 and more precisely in figure 7.4, the more the backlog coefficient is different from the one of the reference case, the bigger the extra costs are. Basically, that happens because the optimal policy for the reference case is used for a scenario more and more different and which has different optimal escalation levels.

7.1.3 Costs of the production plans

The optimal thresholds also depend on the costs of the production plan which are represented by C_1, C_2, C_3 and C_4. The latter depends on many factors such as the normal hourly cost, the overtime cost, the manpower flexibility, etc..

7.1.3.1 Cost of manpower

The cost of manpower can be highly dependent on the production context where the plant is located. Generally, in the more industrialized countries the manpower costs for the company are higher and, as a consequence, the optimal reaction policy will focus more on minimizing the manpower costs.

Figure 7.5 shows how the optimal reaction levels and the corresponding objective function change with respect to that normal hourly salary. In particular, The horizontal axis represents the hourly salary divided by the one of the real case, which is $40 \frac{euro}{hour \cdot worker}$.

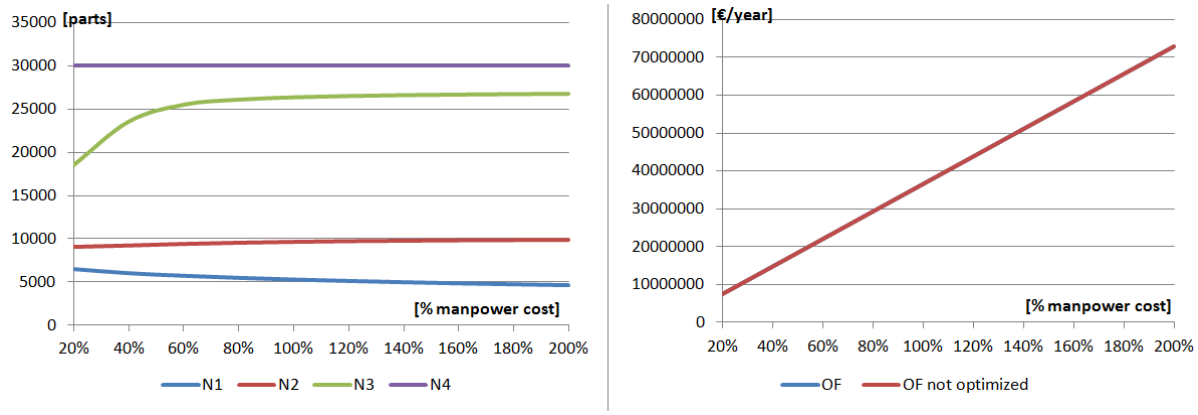

Figure 7.5: Optimal thresholds and OF as a function of the hourly salary

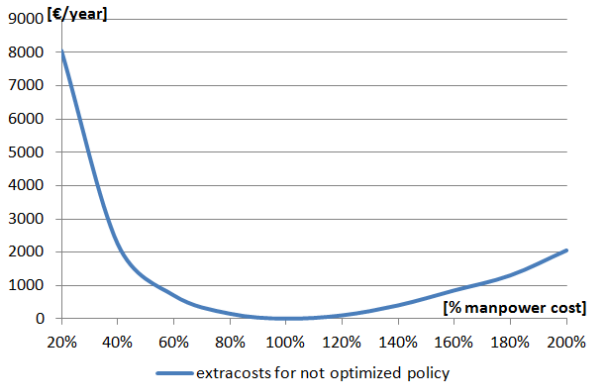

Figure 7.6: Extra costs if the policy is not changed

It can be concluded that, as the manpower costs decrease, the optimal reaction policy becomes more and more reactive because the company must pay more attention on the backlog and inventory costs. In particular, if the manpower costs remain as in the reference case (or even more costly), the priority of the company must be to use it as efficiently as possible. By contrast, if the manpower costs is less than 60% the current value for the studied plant (reference case), the optimization of the inventory costs becomes particularly important compared to manpower costs and the third threshold N_3 decreases quickly.

In this case, it is also worthy to show how much extra cost the company incurs if the optimal policy for the reference case is not changed, when the manpower costs more, and used in other scenarios. As it can be already seen in figure 7.6, the more the manpower costs are different from the one of the reference case, the bigger the extra costs are. Basically, that happens because the optimal policy for the reference case is used for a scenario which is more and more different and, as a consequence has more and more different optimal escalation levels. Indeed, the more the optimal thresholds are different from the ones of the reference case, the more extra costs the company incur if the thresholds are not changed.

7.1.3.2 Overtime costs

The overtime costs can also change depending on the country where the plant is located. In the considered case it costs 50% more, i.e. $60\frac{euro}{hour \cdot worker}$, than the normal salary if it is planned on the first shift on Saturday and it only influences $C_4 = C_3 + 128 workers \cdot 8 hours \cdot 40\frac{euro}{worker \cdot hour} \cdot (1 + \Delta costOvertime)$. In the considered case $\Delta costOvertime = 0.5$ but in figure 7.7 it is shown how the optimal thresholds and the corresponding objective function varies for a $\Delta costOvertime$ with minimum value equal to zero, i.e. the overtime implies the normal shift costs, and 1, when the overtime costs the double as normal shift.

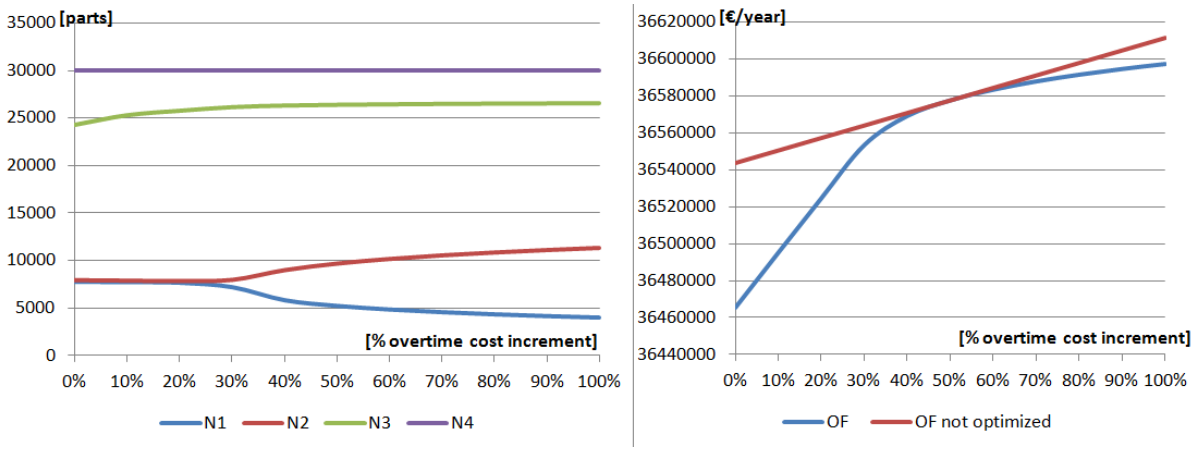

Figure 7.7: Optimal thresholds and OF as a function of the overtime cost

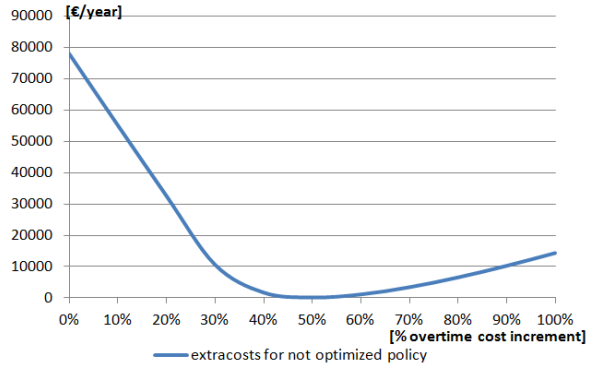

Figure 7.8: Extra costs if the policy is not changed

The mean shift cost for the 15 shifts production plan is $50\frac{euro}{hour \cdot worker}$, i.e. 25% greater than the normal salary, because it includes also the evening and night periods which are paid more. As a consequence, as long as the addition shift in overtime costs less than the 125% of the normal cost, it is suggested to never use the 15 shifts plan because it is less cost-efficient than the one with 16 shifts. On the contrary, if the $\Delta costOvertime$ becomes greater, the 16 shifts plan becomes less cost-efficient and it is used only when the inventory level gets seriously low. The third threshold gets higher for two main reasons. Firstly, since the 16 shifts plan is less used, it is less likely that the inventory level quickly increases and the warehouse gets full and blocks the production line. Secondly, the inventory costs become less important with respect to the ones related to the manpower.

As for the other cost function parameters, it is also interesting to show how much extra cost the company incurs if the reference case optimal policy is not changed, when the overtime costs vary. As it can be already seen in both figure 7.7 and more precisely in figure 7.8, the more the overtime costs are different from the ones of the reference case, the bigger the extra costs are. Basically, when the preferable thresholds are very different from the current ones, the company incurs big extra costs.

7.1.3.3 Manpower flexibility

In the considered case, when the inventory level must be reduced by moving some workers to other departments, it assumed that they can be employed with an efficiency of 60%, which takes into consideration that those workers may not have the proper skills to work somewhere else or they may be employed in departments which are already saturated. The figure 7.9 shows how the optimal thresholds would change if the flexibility would be different.

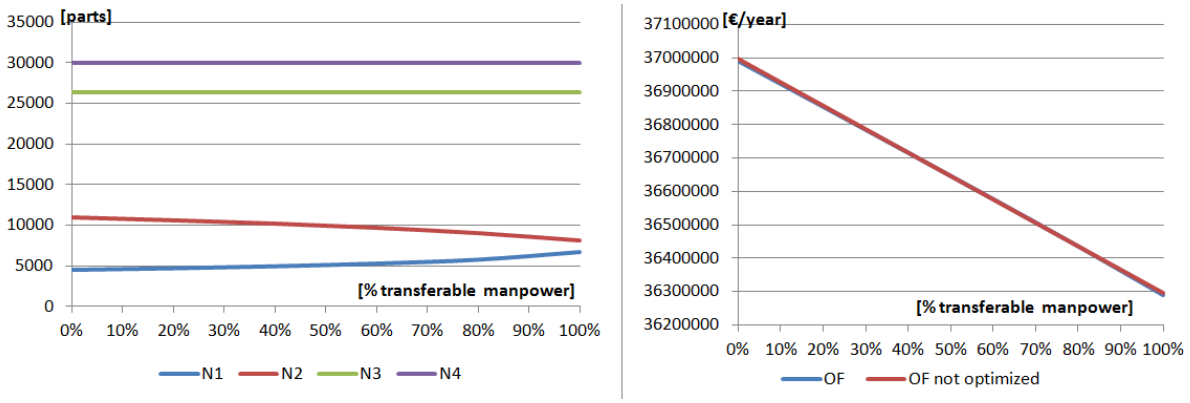

Figure 7.9: Optimal thresholds and OF as a function of the manpower flexibility

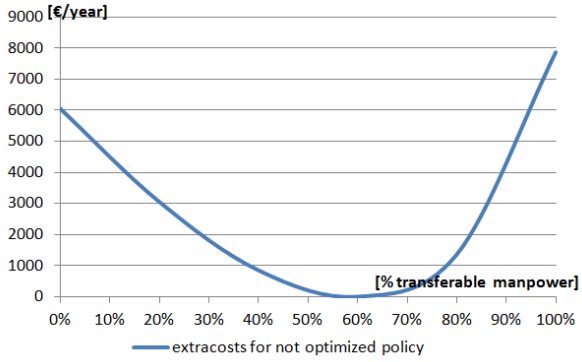

Figure 7.10: Extra costs if the policy is not changed

It can be noticed that, as the flexibility increases, the range related to the removal of 1 shift also increases because it becomes more cost-efficient. On the contrary, although the range associated to the removal of 3 shifts becomes more cost-efficient as well, its efficiency remains lower than the 14 shifts plan and, as a result, it is only used to limit the blocking risk related to a full warehouse. The first threshold N_1 increases because, since the manpower costs become lower, the minimization of the backlog costs becomes more important and it is better to react before, when the inventory level becomes very low.

Also in this case, if the escalation levels are kept constant and the manpower become less or more transferable, the company incurs some extra costs (see figure 7.10). Since the optimal thresholds are not strongly dependent on the manpower flexibility (figure 7.9), the extra costs are small.

7.2 Production plan characteristics

In this section the coefficients of the objective function and the demand characteristics are kept constant but the expected values and/or the variances of the production plans are modified.

7.2.1 Production expected value

In the first experiment the expected value of each plan is modified and it is shown how the optimal reaction levels change as a result. This experiment can be useful for two main situations. Firstly, to asses the advantage of changing the nominal production rate of the plant, for instance, by adding or removing assembly lines in the department. Secondly, in the design phase where the nominal production rates of the plant must be decided considering the forecast demand behavior.

In order to obtain reasonable results, it must be checked that the demand expected value is lower than the bigger expected value among the production plans, otherwise the assumption that the demand is always satisfied is not true. Furthermore, it is also reasonable that the demand is greater than the lower expected value among the production plans, otherwise the warehouse results almost always full and it is always tried to remove as many shifts as possible.

Considering those two aspects, the chosen range for the ΔE is between -6% and $+18\%$ and the trend of the optimal escalation levels and the corresponding objective function is shown in figure 7.11. On the horizontal axis is represented the expected value divided by the one of the real case expressed as a percentage. In this case, the manpower cost of each plan is multiplied by the percentage of E considered in each case. It means that, for example, if it is considered a $\Delta E = +5\%$, it is reasonable to think that the manpower costs will be also 5% bigger. Otherwise, if the plan costs are kept constant, the objective function decreases because less shifts are needed to produce the same quantity and the graph gives trivial information.

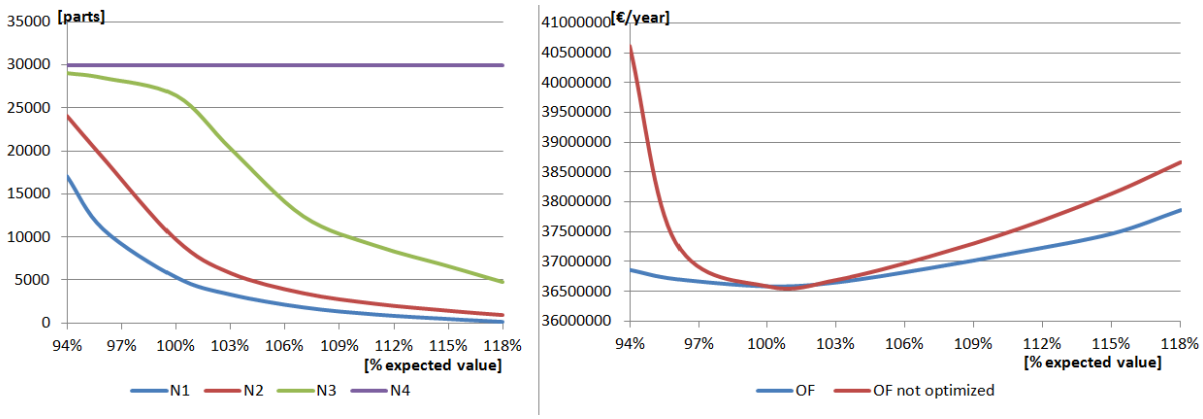

Figure 7.11: Optimal thresholds and normalized OF as a function of the production expected value

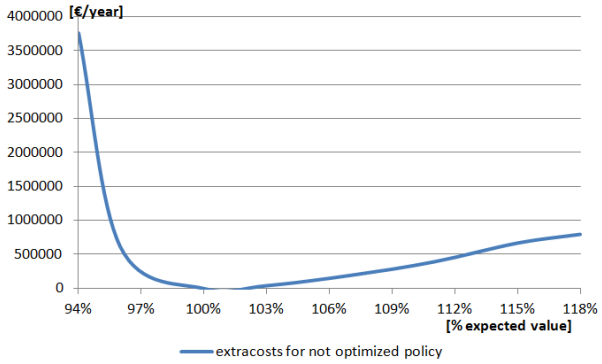

Figure 7.12: Extra costs if the policy is not changed

As the expected value of each plan increases, the plans associated to shift removals become more and more used because less shifts are needed to satisfy the same amount of demand. The objective function trend shows that, since the 14-shift and 15-shift plans are the most cost-efficient, it is better if the demand expected value lies between their expected values. When $\frac{E}{E_{consideredCase}} = 1$, $E_{14shifts} = 12749$ and $E_{15shifts} = 13666$ and, since $E_{demand} = 12990$, this situation is satisfied when $\frac{E}{E_{consideredCase}}$ is between 0.951% and 1.019% and, as it can be seen, it corresponds to the range where the OF has the minimum, which is approximately for $\frac{E}{E_{realCase}} = 1$ and it means that the Bosch production plant has been properly designed to minimize the operative costs for the usual demand.

Figure 7.12 shows the extra costs that the company incurs when the reference case optimal reaction policy is used and never changed even if the production expected value of all the plans changes. It is noticed that a much stronger increment of the extra costs occurs if the plans produce on average less. That happens mainly because, if the reaction

policy is not updated, the backlog costs increase quickly. On the contrary, if the plan expected value increases, it will be more likely to have a full inventory but the costs do not increase as quickly as in the previous case.

7.2.2 Production variance

In this experiment the expected value of each production plan is kept equal as in the reference case but the weekly standard deviation is changed. In particular, it is defined the parameter $\%standardDeviation = \frac{\sigma}{\sigma_{referenceCase}}$ and it varies in the range $[0\%, 500\%]$

For each considered value of the $\%standardDeviation$ and for each plan the μ_nom, p and r are computed as follows:

First of all, it is taken into consideration the available time in the plan and it is multiplied by the $E[\mu_{min} = 2.1337\frac{parts}{min}$ to obtain its weekly expected value $E_{week} = E[\mu_{min}] * T_{available}]$. After that, the real case variance of that plan is considered and multiplied by $\%standardDeviation^2$ to obtain the corresponding variance. Finally, the nominal weekly production volume μ_{nom}, which represents the maximal producible products in a week with that plan, must be computed with the following formula, which imposes that the μ_{nom} coincides always with the same quantile of the distribution.

$\mu_{nom} = E_{week} + (\mu_{nom,realCase} - E_{week}) \cdot \%standardDeviation$

Once those three parameters are known, the p and r rates of the plan can be computed with the method presented in the section "**??**". The results are shown in figure 7.13.

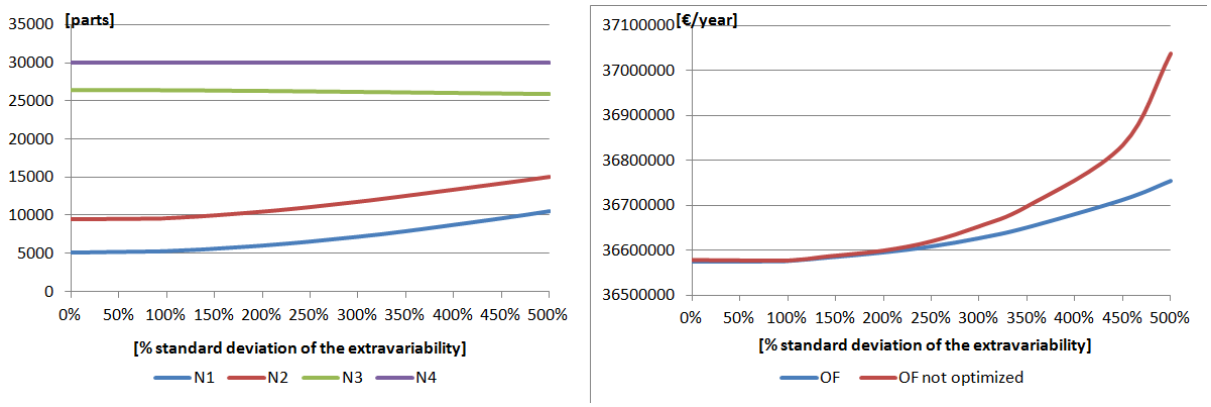

Figure 7.13: Optimal thresholds and OF as a function of the production variance

It can be seen that the more the system increases its variance, the more reactive the optimal escalation policy becomes, namely that, when the inventory level becomes too low or too high, it is tried to implement the reaction policies before. Of course, the variance in the process implies some additional costs which are shown in the objective function graph.

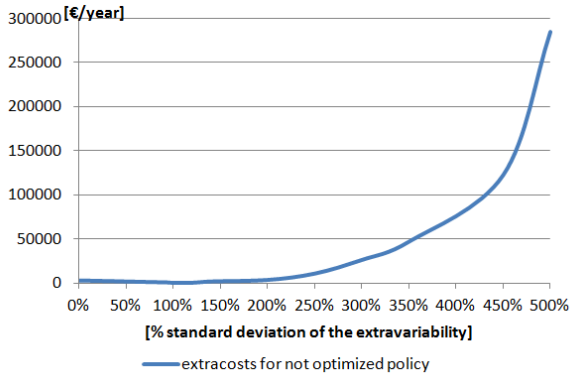

Figure 7.14: Extra costs if the policy is not changed

Figure 7.14 represents the extra costs that the company incurs if the optimized policy for the reference case (i.e. when $\%standardDeviation = 100\%$) is used also when this parameter assumes different values.

7.3 Demand characteristics

In this section the coefficients of the objective function and the plant characteristics are kept constant but the ones of the demand are modified except for the number of demand scenarios and their probabilities.

7.3.1 Demand expected value

It is interesting to understand how the reaction levels along with the operational costs change if the amount of required products per week increases because in a design phase a certain mean demand is assumed but during the whole life of the plant it can change and the reaction policy must be adapted.

In this experiment the parameter $\%demandMeanValue$ is defined and corresponds to $\frac{demandMeanValue}{demandMeanValue_{realCase}}$. Since $demandMeanValue$ must be lower than the maximal expected weekly production quantity among the 4 plans, it can not be greater than $\frac{expectedValue_{4^{th}plan}}{demandMeanValue_{realCase}} = \frac{14584}{12990} = 1.123$. If this condition is not satisfied, the assumption that the demand is never lost is no more valid even if the negative inventory layer is chosen very big. Moreover, it does not make much sense to consider a mean demand lower than the minimal production quantity of the plan with 12 shifts because, in that case, the warehouse would be always full and it would block the assembly line and, as a consequence, the manpower would not be used efficiently. This situation is achieved when $\%demandMeanValue = \frac{expectedValue_{1^{th}plan}}{demandMeanValue_{realCase}} = \frac{10914}{12990} = 0.84$. The chosen range for $\%demandMeanValue$ is $[0.91, 1.07]$.

For each $\%demandMeanValue$ the expected Value of each group can be found by multiplying that coefficient by the expected values of the real case. The parameter α

is kept constant and it implies that also the standard deviation of each group increases proportionally to the expected value. As a result, the p and r rates remain the same and the μ_{nom} is equal to the expected value of the group multiplied by $1 + \alpha$.

The results are shown in figure 7.15.

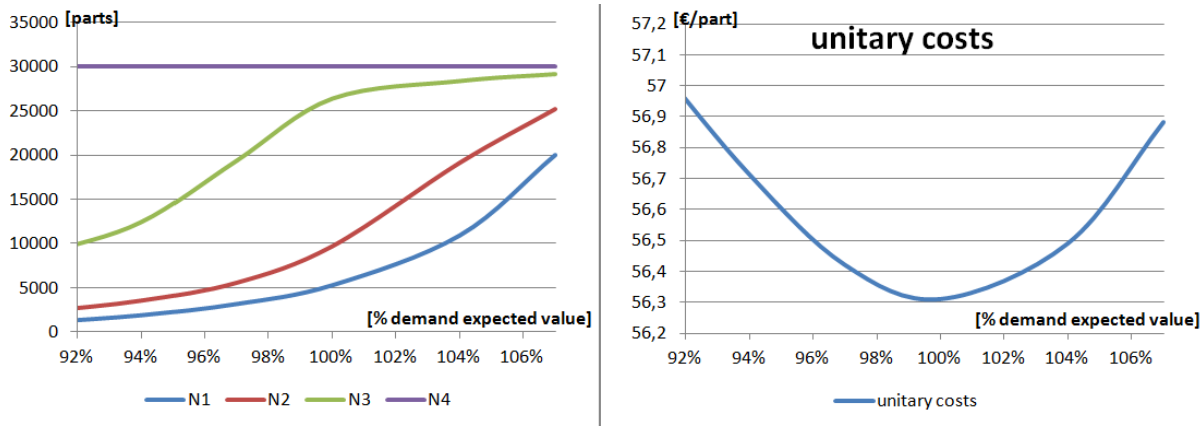

Figure 7.15: Optimal thresholds and unitary cost as a function of the demand expected value

As expected, the graphs show that the higher the demand becomes, the more often the plans with more shifts are used to satisfy the demand. The optimal operational costs increases mainly because more shifts are needed but also the revenues. For that reason, this more useful to represent the cost-efficiency, with which the products are produced, i.e. the unitary cost. Indeed, if the mean weekly demand becomes very different from the weekly production quantity of the most cost-efficient plans, the unitary cost increases.

Also in this case it is possible to plot the trend of the extra costs that the company incurs when only one set of thresholds, which are the result of the optimization in the reference case, are used and the reaction policy is not updated as demand expected value changes. The results are shown in figure 7.16.

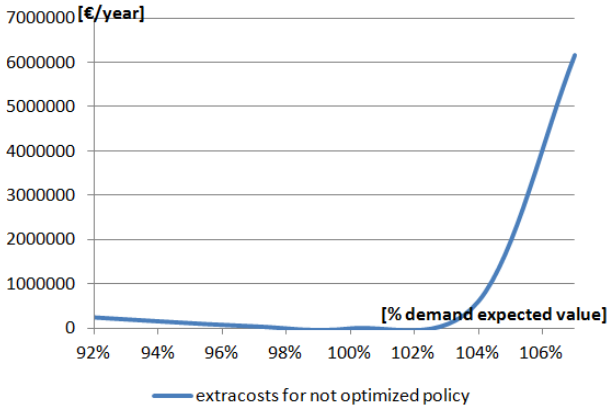

Figure 7.16: Extra costs if the policy is not changed

It is noticed that a much stronger increment of the extra costs occurs if the demand increases its expected value. That happens mainly because, if the reaction policy is not updated, the backlog costs increase quickly. On the contrary, if the demand expected value decreases, it will be more likely to have a full inventory but the costs do not increase as quickly as in the previous case.

7.3.2 Daily demand regularity

The customers generally take products from the final warehouse regularly but some of them only once or twice a week. As a consequence, the demand has an expected variability which must be absorbed by the final buffer. It can be shown how the optimal reaction policy and the corresponding operational costs change as a function of a parameter *demandRange*, which can vary between 0 and 1 and if, for instance, *demandRange*=0.40 as in the real case, it means that the daily demand is always included in the range ±40% with respect to the mean value, which is 12990. Its considered range is [0.1 , 0.9] and, for each value that we want to plot, the expected value of the real case are rescaled as follows:

$$E_1 = 12990 + (E_{1,realCase} - 12990) \cdot \frac{demandRange}{demandRange_{realCase}}$$
$$E_2 = 12990 + (E_{2,realCase} - 12990) \cdot \frac{demandRange}{demandRange_{realCase}}$$
$$E_3 = 12990 + (E_{3,realCase} - 12990) \cdot \frac{demandRange}{demandRange_{realCase}}$$

Basically, foe each demand level the distance from the mean demand value is multiplied by *demandRange*.

Moreover, the parameter α is kept constant and, as a result, the μ_{nom} of each group is obtained by multiplying the expected value by $1 + \alpha$ and the p and r rates are always equal.

The results are shown in figure 7.17.

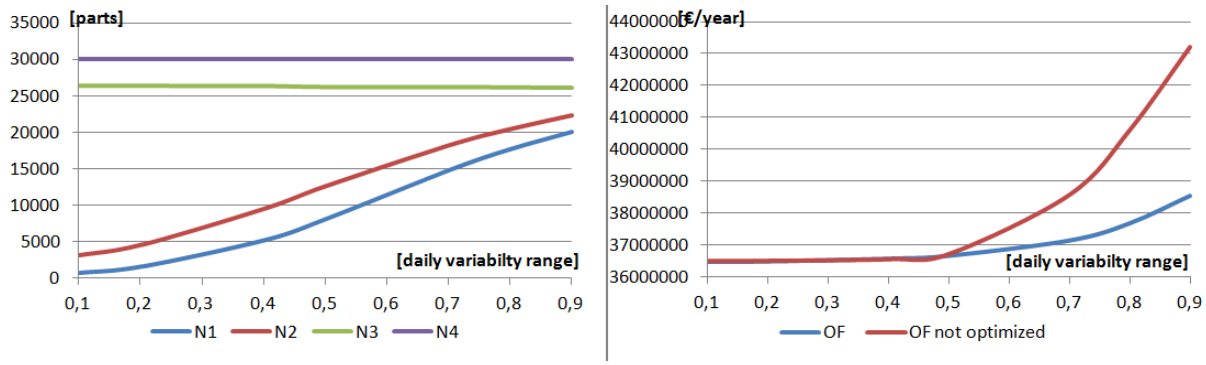

Figure 7.17: Optimal thresholds and OF as a function of the daily demand regularity

It can be seen that the more the planned demand fluctuation increases, the more reactive the optimal escalation policy becomes and, particularly, when the buffer becomes very low because the backlog costs increase much faster than the costs related to the blocking of the production due to a full final warehouse. The objective function graph helps to understand how much an irregular demand would cost to the company and to understand if it is economical to agree with the customer a more constant retrieval plan of final products, for instance, in exchange of a lower price.

Figure 7.18 represents the additional costs for the company if the optimal reaction policy for the reference case ($demandRange$=0.40) is used also when the value of $demandRange$ is different. It can be seen that, if the $demandRange$ increases, the effect of usage of a not-optimized policy is much stronger than in the case where this parameter decreases.

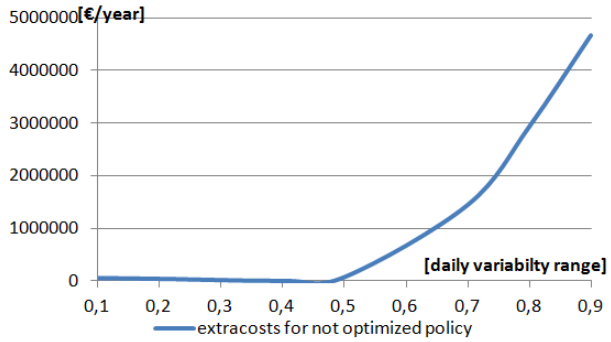

Figure 7.18: Extra costs if the policy is not changed

7.3.3 Demand extra-variance

As already said, the daily demand quantities can also have unexpected variations which are modeled with the fact that each group represents a daily demand which is not

deterministic but randomly distributed. The distribution function is approximated by a sub-Markov chain composed by 1 up and 1 down state and the first one has a production rate equal to the expected value multiplied by $1 + \alpha$.

This experiment is carried out by keeping the same expected values for the three groups of demand and by changing the parameter α in the range $[0, 0.35]$. For each value of it, the formula $1.96 \cdot \sigma_{day} = E_{day} \cdot \alpha$ can be used to find the correspondent weekly variance $VAR_{week} = 5 \cdot \sigma^2_{day} = 5 \cdot (\frac{E_{day} \cdot \alpha}{1.96})^2$.

Once the weekly expected value, variance and production rate of each group and for each chosen α value are known, the corresponding p and r can be found with the method described in the section "**??**" of the Methodology chapter. The results are shown in figure 7.19.

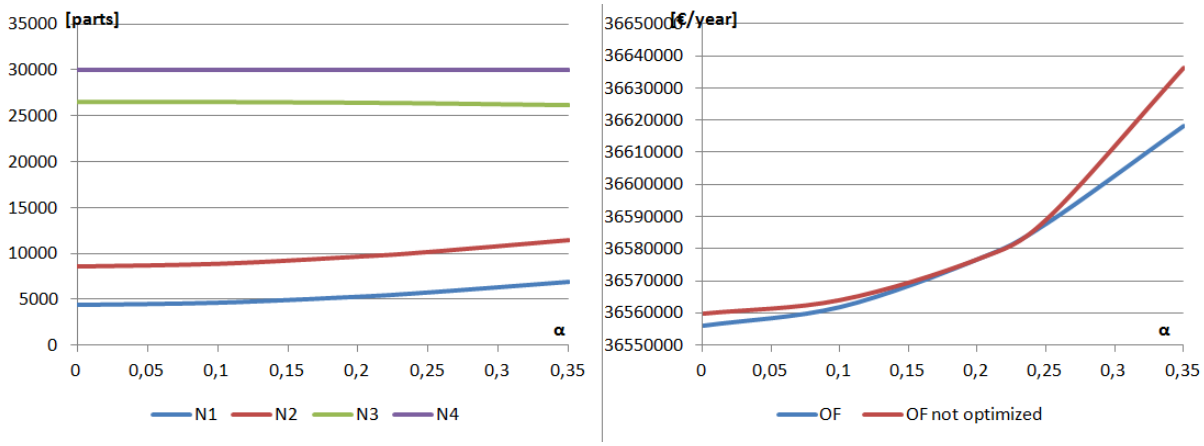

Figure 7.19: Optimal thresholds and OF as a function of the demand extra-variance

The considerations are similar to the ones made for the case where the production variance changes. Indeed, the optimal escalation policy becomes more and more reactive as the variance increases and the first two thresholds change more because the backlog costs would increase very quickly if the reaction policy would not be modified. The second graph is useful to understand how much the unplanned demand variance costs to the company and it is possible to evaluate the convenience to force the customer to respect their scheduled demand, for example, in exchange of a lower price.

For the demand extra-variance it is also possible to plot the extra-costs trend (figure 7.20).

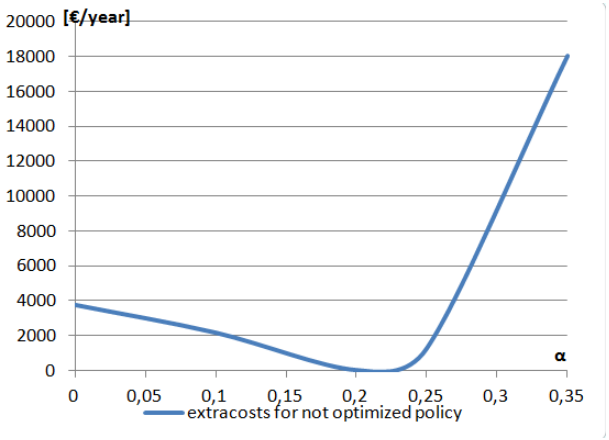

Figure 7.20: Extra costs if the policy is not changed

This graph represents how the extra costs increase if the optimal reaction policy is computed for the reference case and never changed even if the value of α varies. In particular, the extra costs increase quicker if value of α increases since it corresponds to a situation with more backlog risks.

7.4 Maximal inventory level

Since it is noticed that in all the previous cases the optimal upper threshold coincides with the maximal buffer capacity denoted as $maxInv$, it is reasonable to think that, if it could be even bigger, the weekly costs could decrease even more. The figure 7.21 shows how the optimal thresholds change when the parameter $maxInv$ increases.

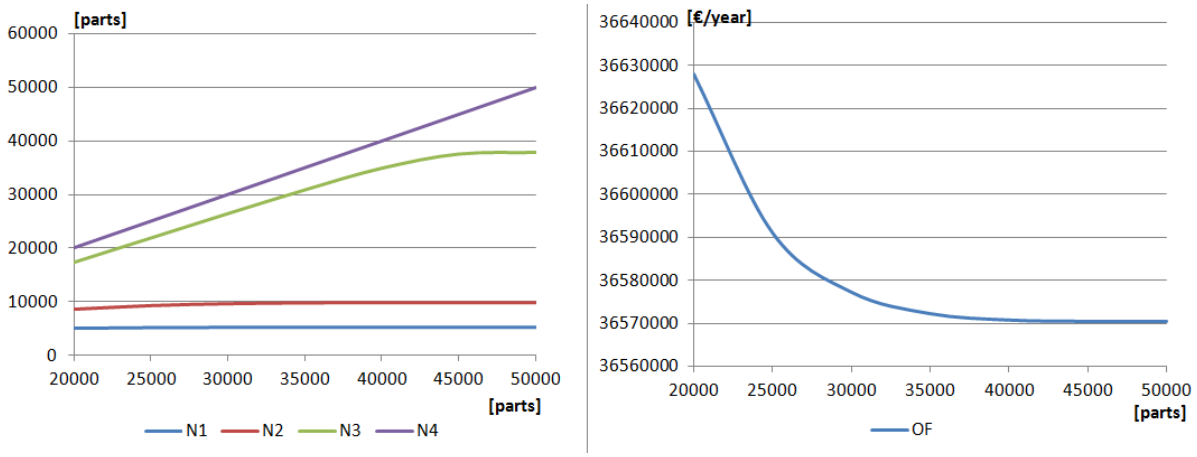

Figure 7.21: Optimal thresholds and OF as a function of the maximal inventory level

Firstly, the upper threshold always coincides to $maxInv$ and it means that it is never economical to stop the production before the warehouse is full. Then, the third threshold increases more or less proportionally with it until $maxInv$ is equal to 40000 and then it tends asymptotically to approximately 38000 parts. It means that, even if the maximal buffer capacity is infinite, it is economical to remove shifts if the inventory level risks to become bigger. The other two thresholds are not strongly affected by the parameter $maxInv$ because, on one hand, the system tries to use the most cost-efficient policy as much as possible and, on the other hand, the system must react to avoid the stockout situation. If the width of the third range is big, those two aspects of the optimal reaction policy are almost independent from the $maxInv$. The graph concerning the objective function is useful, for example, to evaluate the convenience to enlarge the final warehouse. This action implies some investment costs but makes the operational costs of the future periods be lower. For example, the actual $maxInv$ is equal to 30000 parts but, if it was increased to 40000 parts, that would make the company save $130\frac{euro}{week}$ which corresponds to $6500\frac{euro}{year}$. This yearly saving is quite small compared to a typical investment cost required to enlarge the warehouse and, as a result, it can be concluded that the investment is not profitable.

7.5 Number of reaction policies

Currently, the managers choose the production plans for the following 2 weeks between 4 possibilities, i.e. 12, 14, 15 and 16 shifts per week. However, it is interesting to economically evaluate what implies to have less or more possibilities for the weekly production plan. This experiment is performed by considering a number of plans from 2 to 5. Particularly, the one with 4 possibilities corresponds to the real case and the other ones are created by removing the central plans or adding the possibility to schedule 13 shifts which corresponds to the removal of the second and third shift on Friday, which cost respectively $50\frac{euro}{hour \cdot worker}$ and $60\frac{euro}{hour \cdot worker}$. The extreme plans, i.e. the one with the lower and the higher number of shifts, are always included because otherwise the stability of the system is compromised. The first case includes only them and in the next case the most cost-efficient plan (lower product unitary cost) among the remaining ones is added. The following table shows their cost-efficiency.

number of shifts	$\Delta cost$	cost	$T_{available}$	$E_{weeklyProduction}$	unitary cost
12	-135168	632832	5115	10914	57.98
13	-99123	668877	5545	11831	56.53
14	-54067	713933	5975	12749	56.00
15	0	768000	6405	13666	56.20
16	+61440	829440	6835	14584	56.87

The unitary cost, which represents the cost-efficiency of the plan, is also plotted in figure 7.22.

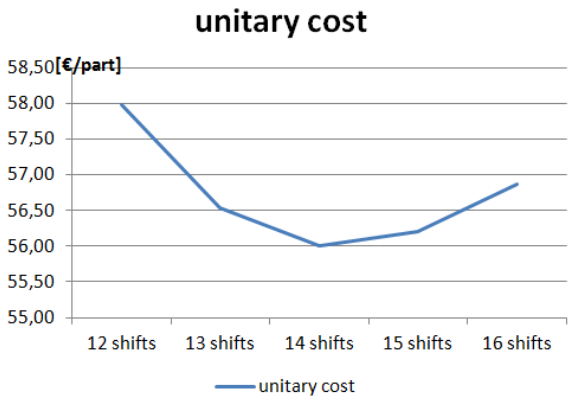

Figure 7.22: Unitary cost of each plan

The following table shows which plans have been chosen for each case.

number of plans	available plans
$2plans$	$12-16shifts$
$3plans$	$12-14-16shifts$
$4plans$	$12-14-15-16shifts$
$5plans$	$12-13-14-15-16shifts$

The results are shown in figure 7.23.

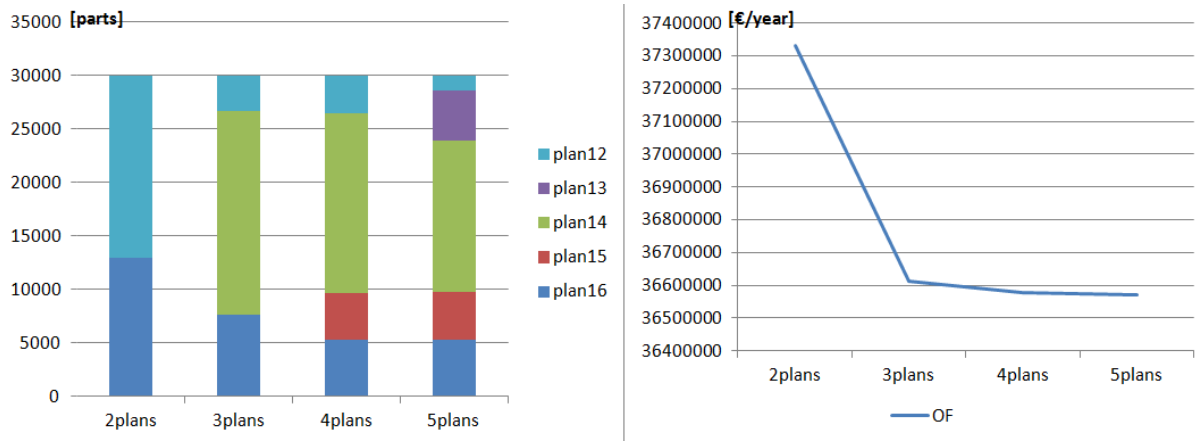

Figure 7.23: Optimal thresholds and OF for each number of available plans

In the case of only two available plans, the operational weekly costs are very big because both plans have a low cost-efficiency. If the one with 14 shifts is added, the operational costs decrease strongly because this plan is the most cost-efficient one and its weekly expected production quantity is almost equal to the mean weekly demand. It

means that, if its range is sufficiently wide, the buffer can absorb the demand variability and the system can almost always produce with the most cost-efficient plan. After that, if the plan with 15 shifts, which also have a great cost-efficiency, is added, it is possible to produce almost always at a low cost with the additional possibility to adjust a bit the weekly production quantities. For that reason, the operational costs decrease even more, although the optimal range of the already added policy is not very large. Finally, if the possibility to choose also the plan with 13 shifts is added to the real case, a saving of $113\frac{euro}{week}$, which corresponds to $5650\frac{euro}{year}$, can be obtained without additional investment costs. As a result, it is suggested to choose between all the 5 plans.

Figure 7.24 shows how the computational time increase, when more plans and, as a consequence, also more ranges are modeled.

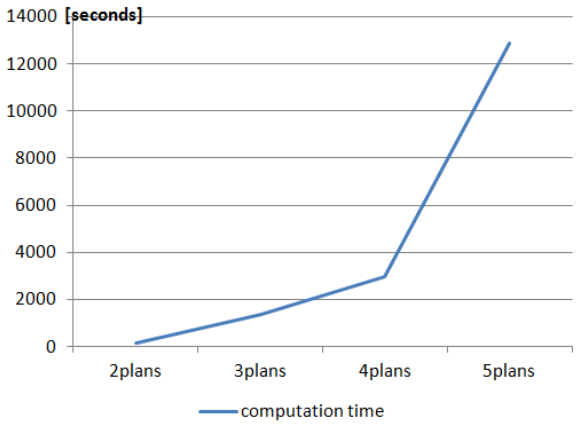

Figure 7.24: Optimal thresholds and OF for each number of available plans

The computational time increases very quickly and it can be seen that by adding the fifth possible plan to the real case, it becomes more than 4 times bigger and it is due to the fact that, on one hand, the computation for each single set of threshold levels becomes more time-consuming because the number of states and roots on each layer and the number of layers increases. On the other hand the optimization must be performed for 5 thresholds instead of 4.

As a consequence, if the required data to model the behavior of the next two weeks are available some hours before the meeting, it is economical to consider also the fifth possibility to further minimize the operational costs. On the contrary, if the time available for the computation is, for example, one hour, the company is forced to consider only 4 possibilities or it is tried to get the information before.

Moreover, the optimal escalation policy for the characteristics of the demand and the production plan of the next two weeks have already been computed in the past, it can be reused without performing a new computation because the result would be exactly the same. The easiest case is the one where the distribution of the weekly production quantity for each plan, the scheduled daily demands and the uncertainty on the daily

demand does not change for many weeks. In this case, the computation is performed only once and the same reaction levels are used until something changes.

7.6 Number of reaction policies - simplified model with biweekly decision

In the previous chapter it has been said that it is possible to model a biweekly production plan decision by considering that every 2 weeks the managers decide the total amount of shifts for both weeks (see section "Biweekly decision delay", chapter 6). In order to have limited number of configuration and thresholds, it is considered that only 3 weekly production plans are available (12,14 and 16 shifts), which corresponds to model 5 possible combined production plans and a decision delay whose expected value is equal to 1 week. The whole modeling has already been done in the previous chapter but now it is investigated how the costs and the optimal thresholds vary if a lower number of policies, which is made vary between 2 and 5, is used. The following table shows their cost-efficiency.

$BiweeklyShifts$	$\Delta cost[\frac{euro}{week}]$	$cost[\frac{euro}{week}]$	$T_{available}[min]$	$E_{weeklyProduction}[part]$	$unitarycost[\frac{euro}{part}]$
24shifts	-135168	632832	5115	10914	57.98
26shifts	-94617	673383	5545	11831	56.92
28shifts	-54067	713933	5975	12749	56.00
30shifts	+3686	771686	6405	13666	56.47
32shifts	+61440	829440	6835	14584	56.87

The unitary cost, which represents the cost-efficiency of the plan, is also plotted in figure 7.25.

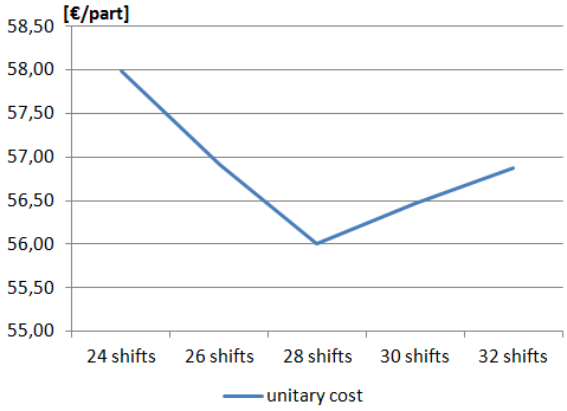

Figure 7.25: Unitary cost of each plan

The results are shown in figure 7.26.

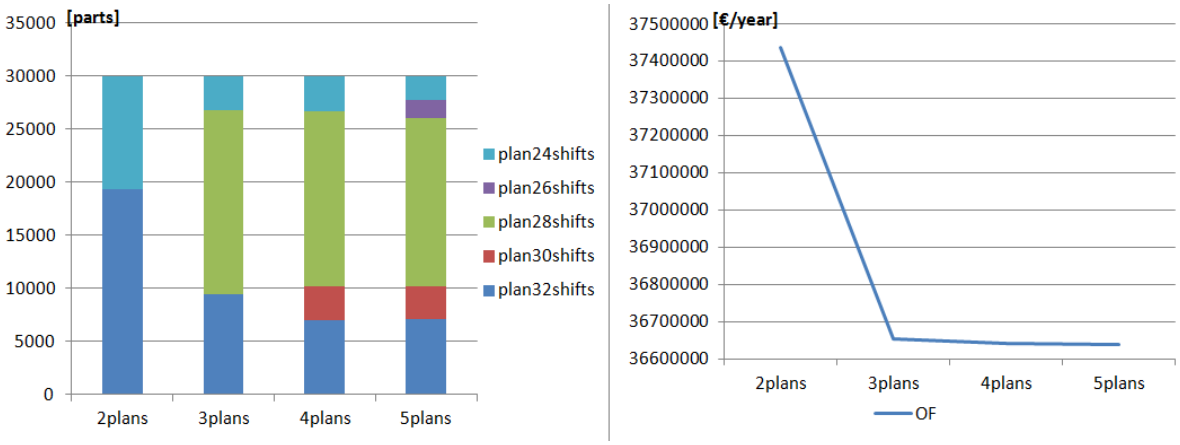

Figure 7.26: Optimal thresholds and OF for each number of available plans

Since the 28-shifts plan is the more cost-efficient and has almost the same expected value as the demand, the total costs can be strongly decreased when it is included. Indeed, the case with 3 plans costs $784100\frac{euro}{year}$ less than the case with only 2 plans. The adding of the 30-shifts plan and, then, of the 26-shifts plan makes the company save a much lower amount of money ($10800\frac{euro}{year}$ and $2200\frac{euro}{year}$ respectively).

8 Conclusions and future research

In this thesis it is shown how the multi-thresholds model presented in [TR13] can be used to model and optimize the production planning and/or the reconfiguration of lean production systems which can be represented by two machines and a buffer between them. The upstream machine can be reconfigured to optimize the operative costs and the downstream machine represents the customer behavior.

The requirements which the system needs for this approach are the followings ones:

- The system can be reconfigured in more than one way
- Only one parameter can be monitored and the system reconfiguration behavior must depend on the instantaneous value of that parameter
- That monitored parameter must be the inventory level
- The user must be able to model the demand behavior with a continuous-time Markov chain

This model can be very useful to solve the production planning problems of the Bosch plant presented in this thesis. In particular, the company issue is to define every two weeks the number of weekly shifts to be planned considering the current inventory level and the plant and demand behavior along with its expected and unexpected uncertainties. The problem is currently solved by defining some fixed ranges for the inventory level. A different number of weekly shifts to be planned is associated to each range and, when the biweekly meeting occurs, the number of shifts for the next 2 weeks are identified considering the current inventory level.

That kind of problem is very common in lean production contexts because, on one hand, the company would like to ensure a service level close to 100% by maintaining a high inventory level but, on the other hand, the company would like to be "lean" and limit the stocking costs and lead times by keeping the inventory level as low as possible. At the same time the manpower must be used as efficiently as possible and, as a consequence, it can be preferable to make the buffer absorb the normal demand fluctuations instead of removing or adding shifts to follow the demand trend.

8.1 Work summary

The first part of this thesis has been dedicated to the presentation of typical problems and solutions in the lean production systems, namely the uncertainties along the

whole supply chain and the possibility to adjust some of its characteristics when some monitored parameters assume dangerous values, for example, when the buffer level gets dangerously low or when the bottleneck of the system is no more able to satisfy the current demand.

After a general reasoning without referring to a specific case, we focus on the subset of problems which includes also the presented Bosch real case. The considered system is composed by two machines, which generally represent both the production plant (upstream machine) and the demand(downstream machine), and a buffer between them. The upstream machine behaves differently depending on the current inventory level and, precisely, for each buffer level a preferable configuration exists and the system will be reconfigured as soon as possible if the current configuration is not the preferable one. This reconfiguration mechanism, if properly designed, is used to limit the impact of risk and uncertainties in both the production plant and demand side.

Thirdly, the real case is presented and, particularly, it is explained which products are produced, the layout of the plant, the used Kanban-policy, the production planning policy and the customer behavior. It is clarified how the system characteristics, which are needed for the modeling, are determined starting from the available data, how they are modeled and how the outputs are used to determine the corresponding operative costs with the defined cost function. After that, it is shown the economical advantage of optimizing the escalation thresholds by considering explicitly the non-differential costs.

The last chapter is entirely dedicated to the considerations about how the optimal escalation levels vary if some characteristics of both the production plant and demand side change. This final parts is particularly important from the company point of view because it has production plants in lots of different countries, where the different cost coefficients, plant characteristics, manpower flexibility and customer behavior can be strongly different. Moreover, it is quantified the advantage of adapting the reaction thresholds to the current production environment. In particular, it is shown the extra costs for the company if a not optimized policy is used. As a consequence, those results provide important guidelines about how the reaction policy must be change and how important is to change it.

8.2 Research contributions

The main breakthroughs of my thesis are the following ones:

- This work defines the general approach to deal with the optimization of the reaction policy defined by any number of combinations of adjustable parameters and any number of monitored indexes. Once the number of ranges for each parameter is chosen, an objective function can be defined and optimized by finding the best position for each threshold. This approach can be used to deal with a huge variety of problems but in this thesis it is only introduced.

- The modeling presented in chapter 4 is based on the analytical tool presented in [TR13] and represents a general methodology to deal with reactive systems which

can be modeled with an upstream stage, which stores the produced parts in a buffer and can be reconfigured to face impending risks, and a downstream stage, which retrieves parts from the buffer with a defined behavior. The risks which can be identified are only related to the amount of products in the buffer.

- The developed analytical model is able to provide a fast and exact solution for the computation of the performances of complex systems.

- Furthermore, it is capable to model risks and uncertainties coming from the different stages of the supply chain. For instance, if the upstream machine represents the production plant of the company and the downstream machine represents the customer, it is possible to model not only the production and customer variability but also the component shortage, due to an unreliable supplier, with a pseudo-failure in the upstream machine.

- The developed tool can also consider different levels of reaction activities (operational, tactical and strategic) by modeling different implementation delays and considering different costs for each configuration.

- The model can also be used in the plant design phase when the nominal production capacity must be chosen and the demand behavior is assigned. It is general decided to have almost the same production rate for the demand and the most cost-efficient production plans in order to make them able to satisfy a regular demand and to use the other plans to avoid extra stockout or inventory costs.

8.3 Considerations for the company

The following considerations about the results for the Bosch case can be done:

- The use of a quantitative model, instead of methods and formulas which do not consider directly the related costs, represents for Bosch a big opportunity to be more cost-efficient and to become even more "lean". Indeed, it is suggested to compute the optimal escalation levels just before the meeting and, precisely, to model the forecast system behavior of the following two weeks. Then, once the optimal escalation levels are computed, the current inventory level identifies the decision to be taken. This procedure assures that the reaction policy is updated if some characteristics of the supply chain change

- The results obtained with this model underline that the currently used escalation ranges are very different from the optimal ones and, precisely, the current reaction policy is too reactive. Indeed, it would be better use the 14-shifts plan for a wider buffer range because it is the most cost-efficient and its expected value is equal to the one of the demand and react with different production plans only if the inventory level becomes very high or very low.

- Currently, the meeting, during which the number of shifts is decided, is planned every 2 weeks. However, it is strongly recommended to consider the possibility to evaluate again the inventory level at the end of the first week and, if necessary, to reschedule the number of shifts also at the beginning of the second week. Although this method provides a shorter planning interval for the upstream departments, it has the big advantage of choosing the number of shifts for the second week without uncertainty.

- It is not strictly necessary to compute the optimal escalation thresholds every time but it is possible to build graphs like the ones presented in the chapter 7, where only one parameter is varied, and save them. We can use them again when similar conditions for the next two weeks occur. For instance, it can happen that a graph can be reused if all the characteristics of both the plant and the demand remain equal to the reference case except for the mean demand value. In this case, it is, for instance, possible to reuse the graph 7.15 without performing the computation again.

- Since it has been found that it is better if the production is never stopped by the absence of Kanbans until the final warehouse becomes full, it is also suggested to define the total number of Kanbans depending on the level of the upper threshold N_4, which in the considered case always coincides to the total buffer capacity, and not vice versa. Since the absence of production Kanbans at the beginning of line is already experienced when the %80 is attached to the parts, the proper number of Kanbans can be computed as follow to assure that the inventory level can actually reach N_4:
  $0.8 \cdot NK = \frac{N_4}{NPK}$

- As shown in chapter 7 with some examples, it can be generally stated that the more the uncertainties and variability in the production environment increase, the more reactive and costly the optimal escalation policy becomes. The company can assess the convenience of implementing some changes in the supply chain by comparing their cost with the economical advantage which that action implies. For example, it has been shown that if the scheduled demand would be equal to real one, i.e. the customers strictly comply with the scheduled demand, the operative costs can be reduced. Therefore it is possible to offer the costumers a lower price in exchange of a more predictable real demand behavior defined by contract. The same reasoning can be done with the normal demand fluctuation by making the scheduled daily demand more constant, with the production plant uncertainties by decreasing the daily production variability and so on.

- It is suggested to add the possibility of planning also 13 shifts in a week (5 possible plans instead of 4) because our model estimates a saving of $5600 \frac{euro}{year}$ by simply adding this possibility

- It is shown that the upper threshold is always equal to the maximum buffer capacity, even in the experiment where the buffer capacity is almost doubled. It means

that, if a certain buffer capacity is available and it can't be used otherwise, it is never economical to block the production before the buffer is full. As a result, for this particular case, it is possible to set the upper threshold equal to the maximum inventory level and optimize only 3 thresholds instead of 4. That implies a much lower computational speed as in the performed experiments.

8.4 Future developments

Some aspect which can be developed in future are:

- In many real applications the monitored parameters can be more than one and/or different from the inventory levels of the final products. In that case our model can be no more used but other analytical models can be developed to study them.
- The delay is currently exponentially distributed and it represents probably one of the strongest approximation because the exponential distribution implies a big variance and, as a result, the optimization returns a reaction policy which is more reactive than the real optimal one. Indeed, if the variability on the future increases and, as a result, the risks, the model will consider a narrower buffer range as "safe" and will react sooner if the buffer dangerously increases or decreases. This aspect could be improved to make the model more accurate.
- The time to transition among the demand scenarios is exponentially distributed as well and it also represents one of the strongest approximations because, as already said in the previous point, the exponential distribution implies a big variance and, as a result, the optimization returns a reaction policy which is more reactive than the optimal one in the reality. This aspect could be also improved in improve the model accuracy.
- In the considered real case we have considered only two machines which which represent the final part of the supply chain (final production department and customers) and the unreliability of the previous parts is considered implicitly in the upstream machine because, when the components are missing, it is equivalent to a pseudo-failure, which causes performance losses. However, if the failure data provide only information about the real failures, it can be useful to develop a decomposition method.
- Currently, the backlog situation is modeled with a negative buffer range which is chosen wide enough to make the probability of lying on the lower threshold very small. Since it must be always checked that this probability is negligible, it could be better to modify the equation in [TR13] to model an infinite negative buffer.
- The developed Matlab code uses a multiprecision toolbox, which allows to deal with the exponent range of quadruple precision variables (between 10^{-4965} and 10^{+4932}). Indeed, as the system complexity and buffer size increase, the overflow

problems using the exponent range of the double variables become more and more likely. However, that slows down the computational speed because Matlab must compute with "multiprecision objects" instead of double variables. It has been experienced that without using the toolbox the computational speed can be much higher or the model can be built with more states for a better production distribution approximation and/or more thresholds if needed. In order to do that some other programming languages, which allow the use of quadruple precision variables, like c++ can be used.

Considering all those aspects, the approach presented to study the reactive manufacturing systems can be further studied and includes plenty of further development possibility, which can be studied in the following years.

Bibliography

[ACM14] Ramiz Assaf, Marcello Colledani, and Andrea Matta. Analytical evaluation of the output variability in production systems with general markovian structure. *OR Spectrum*, 36(3):799–835, 2014.

[AE99] MS Akturk and F Erhun. An overview of design and operational issues of kanban systems. *International Journal of Production Research*, 37(17):3859–3881, 1999.

[AG07] Ronald G Askin and Jeffrey B Goldberg. *Design and analysis of lean production systems*. John Wiley & Sons, 2007.

[AGMG93] Ronald G Askin, M GEORGE MITWASI, and Jeffrey B Goldberg. Determining the number of kanbans in multi-item just-in-time systems. *IIE transactions*, 25(1):89–98, 1993.

[AHC14] A. Angius, A. Horváth, and M. Colledani. Moments of accumulated reward and completion time in markovian models with application to unreliable manufacturing systems. *Performance Evaluation*, 75–76(0):69 – 88, 2014.

[Axs07] Sven Axsaeter. *Inventory control*, volume 90. Springer, 2007.

[Bal92] R. H. Ballou. *Business logics management*. 1992.

[BC87] Gabriel R Bitran and Li Chang. A mathematical programming approach to a deterministic kanban system. *Management Science*, 33(4):427–441, 1987.

[BCG97] Asbjoern M Bonvik, CE Couch, and Stanley B Gershwin. A comparison of production-line control mechanisms. *International journal of production research*, 35(3):789–804, 1997.

[BDMF01] Bruno Baynat, Yves Dallery, Maria Di Mascolo, and Yannick Frein. A multi-class approximation technique for the analysis of kanban-like control systems. *International Journal of production research*, 39(2):307–328, 2001.

[Ber92] Blair J Berkley. A review of the kanban production control research literature. *Production and operations management*, 1(4):393–411, 1992.

[Boo05] Karin Boonlertvanich. Extended-conwip-kanban system: control and performance analysis. 2005.

[BRA94] ARMANDO BRANDOLESE. The problems of total quality. *Production planning & Control*, 5(4):330–336, 1994.

[BTS92] Emilio Bartezzaghi, Francesco Turco, and Gianluca Spina. The impact of the just-in-time approach on production system performance: a survey of italian industry. *International Journal of Operations & Production Management*, 12(1):5–17, 1992.

[Buz89] John A Buzacott. Queueing models of kanban and mrp controlled production systems. *Engineering Costs and Production Economics*, 17(1):3–20, 1989.

[CCS04] Raffaella Cagliano, Federico Caniato, and Gianluca Spina. Lean, agile and traditional supply: how do they impact manufacturing performance? *Journal of Purchasing and Supply Management*, 10(4):151–164, 2004.

[CFT+10] L Cristaldi, M Faifer, S Toscani, A Ferrero, S Ferrari, M Lazzaroni, M Garetti, S Ierace, and S Cavalier. Tecniche di diagnosi basate sul'analisi della firma elettrica. In *Congresso Nazionale Gruppo Misure Elettriche ed Elettroniche*, pages 367–376. Edizioni Università di Cassino, 2010.

[CG10] Seok Ho Chang and Stanley B Gershwin. Modeling and analysis of two unreliable batch machines with a finite buffer in between. *IIE Transactions*, 42(6):405–421, 2010.

[CGT08] A Cannata, M Gerosa, and M Taisch. Socrades: A framework for developing intelligent systems in manufacturing. In *Industrial Engineering and Engineering Management, 2008. IEEM 2008. IEEE International Conference on*, pages 1904–1908. IEEE, 2008.

[CKT09] Alessandro Cannata, Stamatis Karnouskos, and Marco Taisch. Energy efficiency driven process analysis and optimization in discrete manufacturing. In *Industrial Electronics, 2009. IECON'09. 35th Annual Conference of IEEE*, pages 4449–4454. IEEE, 2009.

[CL05] Gerard P Cachon and Martin A Lariviere. Supply chain coordination with revenue-sharing contracts: strengths and limitations. *Management science*, 51(1):30–44, 2005.

[CMMT05] E Cagno, F Magalini, G Micheli, and P Trucco. Supporting lean production through waste minimization: The unseen costs issue. 2005.

[CMT10] M. Colledani, A. Matta, and T. Tolio. Analysis of the production variability in multi-stage manufacturing systems. {*CIRP*} *Annals - Manufacturing Technology*, 59(1):449 – 452, 2010.

[CS60] Andrew J Clark and Herbert Scarf. Optimal policies for a multi-echelon inventory problem. *Management science*, 6(4):475–490, 1960.

[CS00] Maria Caridi and Andrea Sianesi. Multi-agent systems in production planning and control: An application to the scheduling of mixed-model assembly lines. *International Journal of Production Economics*, 68(1):29–42, 2000.

[CS04] Sunil Chopra and ManMohan S Sodhi. Supply-chain breakdown. *MIT Sloan management review*, 2004.

[DFDM00] C Duri, Y Frein, and M Di Mascolo. Comparison among three pull control policies: Kanban, base stock, and generalized kanban. *Annals of Operations Research*, 93(1-4):41–69, 2000.

[DG92] Yves Dallery and Stanley B Gershwin. Manufacturing flow line systems: a review of models and analytical results. *Queueing systems*, 12(1-2):3–94, 1992.

[DHMMO89] Jean-Luc Deleersnyder, Thom J Hodgson, Henri Muller-Malek, and Peter J O'Grady. Kanban controlled pull systems: an analytic approach. *Management Science*, 35(9):1079–1091, 1989.

[DL00] Yves Dallery and George Liberopoulos. Extended kanban control system: combining kanban and base stock. *Iie Transactions*, 32(4):369–386, 2000.

[FDMD95] Yannick Frein, Maria Di Mascolo, and Yves Dallery. On the design of generalized kanban control systems. *International Journal of Operations & Production Management*, 15(9):158–184, 1995.

[FSHK03] T Freiheit, M Shpitalni, S Jack Hu, and Y Koren. Designing productive manufacturing systems without buffers. *CIRP Annals-Manufacturing Technology*, 52(1):105–108, 2003.

[Ger00] Stanley B Gershwin. Design and operation of manufacturing systems: the control-point policy. *IIE transactions*, 32(10):891–906, 2000.

[GFF07] Stanley B Gershwin and Saeideh Fallah-Fini. A general model and analysis of a discrete two-machine production line. *Analysis of Manufacturing Systems*, 2007, 2007.

[GGF+09] Elisa Gebennini, Andrea Grassi, Cesare Fantuzzi, Stanley B Gershwin, and Irvin C Schick. On the introduction of a restart policy in the two-machine, one-buffer transfer line model. In *Proceedings of the 7th international conference on stochastic models of manufacturing and service operations*, pages 81–88, 2009.

[GGF+13] Elisa Gebennini, Andrea Grassi, Cesare Fantuzzi, Stanley B Gershwin, and Irvin Cemil Schick. Discrete time model for two-machine one-buffer transfer lines with restart policy. *Annals of Operations Research*, 209(1):41–65, 2013.

[GH05] John Geraghty and Cathal Heavey. A review and comparison of hybrid and pull-type production control strategies. *OR Spectrum*, 27(2-3):435–457, 2005.

[GMM+01] Stanley B Gershwin, Nicola Maggio, A Matta, T Tolio, and L Werner. Analysis of loop networks by decomposition. In *Third Aegean International Conference on Analysis and Modeling of Manufacturing Systems*, pages 239–248, 2001.

[Gro93] Harry Groenevelt. The just-in-time system. *Handbooks in Operations Research and Management Science*, 4:629–670, 1993.

[GT03] Stephen C Graves and Brian T Tomlin. Process flexibility in supply chains. *Management Science*, 49(7):907–919, 2003.

[GW07] Stanley B Gershwin and Loren M Werner. An approximate analytical method for evaluating the performance of closed-loop flow systems with unreliable machines and finite buffers. *International Journal of Production Research*, 45(14):3085–3111, 2007.

[HIX10] Wallace J Hopp, Seyed MR Iravani, and Wendy Lu Xu. Vertical flexibility in supply chains. *Management Science*, 56(3):495–502, 2010.

[HK96] Chun-Che Huang and Andrew Kusiak. Overview of kanban systems. 1996.

[Hon05] KKB Hon. Performance and evaluation of manufacturing systems. *CIRP Annals-Manufacturing Technology*, 54(2):139–154, 2005.

[HRT83] Philip Y Huang, Loren P Rees, and Bernard W Taylor. a simulation analysis of the japanese just-in-time technique (with kanbans) for a multiline, multistage production system. *Decision Sciences*, 14(3):326–344, 1983.

[HT02] András Horváth and Miklós Telek. Phfit: A general phase-type fitting tool. In *Computer Performance Evaluation: Modelling Techniques and Tools*, pages 82–91. Springer, 2002.

[HWL07] Xin-Feng He, Su Wu, and Quan-Lin Li. Production variability of production lines. *International Journal of Production Economics*, 107(1):78–87, 2007.

[JG95] William C Jordan and Stephen C Graves. Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4):577–594, 1995.

[KD00] Fikri Karaesmen and Yves Dallery. A performance comparison of pull type control mechanisms for multi-stage manufacturing. *International Journal of production economics*, 68(1):59–71, 2000.

[KG83] Joseph Kimemia and Stanley B Gershwin. An algorithm for the computer control of a flexible manufacturing system. *AIIE Transactions*, 15(4):353–362, 1983.

[KHW98] Yoram Koren, S Jack Hu, and Thomas W Weber. Impact of manufacturing system configuration on performance. *CIRP Annals-Manufacturing Technology*, 47(1):369–372, 1998.

[Kim88] George E Kimball. General principles of inventory control. *Journal of manufacturing and operations management*, 1(1):119–130, 1988.

[KKRW87] Lee J Krajewski, Barry E King, Larry P Ritzman, and Danny S Wong. Kanban, mrp, and shaping the manufacturing environment. *Management science*, 33(1):39–57, 1987.

[KP07] C Sendil Kumar and R Panneerselvam. Literature review of jit-kanban system. *The International Journal of Advanced Manufacturing Technology*, 32(3-4):393–408, 2007.

[Kra88] John F Krafcik. Triumph of the lean production system. *Sloan management review*, 30(1):41–51, 1988.

[KS05] Paul R Kleindorfer and Germaine H Saad. Managing disruption risks in supply chains. *Production and operations management*, 14(1):53–68, 2005.

[LD00] George Liberopoulos and Yves Dallery. A unified framework for pull control mechanisms in multi-stage manufacturing systems. *Annals of Operations Research*, 93(1-4):325–355, 2000.

[Lee04] Hau L Lee. The triple-a supply chain. *Harvard business review*, 82(10):102–113, 2004.

[Li13] Z. Li. *Design and Analysis of Robust Kanban System in an Uncertain Environment*. PhD thesis, 2013.

[Lim09] Michael K Lim. *Supply chain network design in the presence of disruption risks*. PhD thesis, NORTHWESTERN UNIVERSITY, 2009.

[LMT03] R. Levantesi, A. Matta, and T. Tolio. Performance evaluation of continuous production lines with machines having different processing times and multiple failure modes. *Performance Evaluation*, 51(2–4):247 – 268, 2003. Queueing Networks with Blocking.

[LPW04] Hau L Lee, Venkata Padmanabhan, and Seungjin Whang. Information distortion in a supply chain: the bullwhip effect. *Management science*, 50(12_supplement):1875–1886, 2004.

[Mag00] Nicola Maggio. An analytical method for evaluating the performance of closed loop production lines with unreliable machines and finite buffer. *Master's thesis, Politecnico di Milano*, 2000.

[MdASL05] Victor Martinez de Albeniz and David Simchi-Levi. A portfolio approach to procurement contracts. *Production and Operations Management*, 14(1):90–114, 2005.

[MFG+11] MARCO MACCHI, LUCA FUMAGALLI, MARCO GARETTI, EMANUELE DOVERE, SERGIO CAVALIERI, STEFANO IERACE, LOREDANA CRISTALDI, MARCO ROSSI, R Zaltieri, and M Vernieri. Metodologia di applicazione della firma elettrica. *Manutenzione Tecnica e Management*, 18(7/8):12–13, 2011.

[MGG10] Federico Mauri, Marco Garetti, and Alessandro Gandelli. A structured approach to process improvement in manufacturing systems. *Production Planning & Control*, 21(7):695–717, 2010.

[MMGT09] N Maggio, A Matta, SB Gershwin, and T Tolio. A decomposition approximation for three-machine closed-loop production systems with unreliable machines, finite buffers and a fixed population. *IIE Transactions*, 41(6):562–574, 2009.

[Mon83] Yasuhiro Monden. *Toyota production system: practical approach to production management*. Industrial Engineering and Management Press, Institute of Industrial Engineers Norcross, GA, 1983.

[Mul14] Multiprecision toolbox for matlab, 2014.

[NvCFF05] Peter Nyhuis, Gregor von Cieminski, Andreas Fischer, and Klaus Feldmann. Applying simulation and analytical models for logistic performance prediction. *CIRP Annals-Manufacturing Technology*, 54(1):417–422, 2005.

[Ohn88] Taiichi Ohno. *Toyota production system: beyond large-scale production*. Productivity press, 1988.

[PASB10] Margherita Pero, Nizar Abdelkafi, Andrea Sianesi, and Thorsten Blecker. A framework for the alignment of new product development and supply chains. *Supply Chain Management: An International Journal*, 15(2):115–128, 2010.

[PB96] Marco Perona and Cristiano Benucci. The integrated kanban system: a new software tool for kanban production. In *IT and Manufacturing Partnerships: Delivering the Promise: Proceedings of the Conference on Integration in Manufacturing, Galway, Ireland, 2-4 October 1996*, volume 7, page 75. Ios PressInc, 1996.

[PJT13] Vittaldas V Prabhu, Hyun Woo Jeon, and Marco Taisch. Simulation modelling of energy dynamics in discrete manufacturing systems. In *Service Orientation in Holonic and Multi Agent Manufacturing and Robotics*, pages 293–311. Springer, 2013.

[PRTIH87] Patrick R Philipoom, Loren P Rees, BERNARD W TAYLOR III, and Phlip Y Huang. An investigation of the factors influencing the number of kanbans required in the implementation of the jit technique with kanbans. *International Journal of Production Research*, 25(3):457–472, 1987.

[PST10] Alberto Portioli Staudacher and Marco Tantardini. Lean production implementation: a survey in italy. *Dirección y Organización*, (35):52–60, 2010.

[RKTT11a] MONICA ROSSI, ENDRIS KERGA, MARCO TAISCH, and SERGIO TERZI. Lean product development: Fact finding research in italy. In *IESM 2011. International Conference on Industrial Engineering and Systems Management*. FR, 2011.

[RKTT11b] MONICA ROSSI, ENDRIS KERGA, MARCO TAISCH, and SERGIO TERZI. Proposal of a reference method for identification of wastes in new product development process. In *XVI Summer School" Francesco Turco"-2011." BREAKING DOWN THE BARRIERS BETWEEN RESEARCH AND INDUSTRY"*. Università di Padova, 2011.

[SB11] Andrea Sianesi and Alessandro Brun. Just in time, lean production e six sigma per migliorare il sistema di produzione. 2011.

[SD06] Lawrence V Snyder and Mark S Daskin. Stochastic p-robust location problems. *IIE Transactions*, 38(11):971–985, 2006.

[SG09] Chuan Shi and Stanley B Gershwin. An efficient buffer design algorithm for production line profit maximization. *International Journal of Production Economics*, 122(2):725–740, 2009.

[Spi90] G. Spina. *I fattori di contesto e l'adozione del JIT*. E. Bartezzaghi, F. Turco, 1990.

[SW00] Günter Schmidt and Wilbert E Wilhelm. Strategic, tactical and operational decisions in multi-national logistics networks: a review and discussion of modelling issues. *International Journal of Production Research*, 38(7):1501–1523, 2000.

[SWH90] Mark L Spearman, David L Woodruff, and Wallace J Hopp. Conwip: a pull alternative to kanban. *The International Journal of Production Research*, 28(5):879–894, 1990.

[Tak03] Katsuhiko Takahashi. Comparing reactive kanban systems. *International Journal of Production Research*, 41(18):4317–4337, 2003.

[Tan00] Barış Tan. Asymptotic variance rate of the output in production lines with finite buffers. *Annals of Operations Research*, 93(1-4):385–403, 2000.

[Tan06a] Christopher S Tang. Perspectives in supply chain risk management. *International Journal of Production Economics*, 103(2):451–488, 2006.

[Tan06b] Christopher S Tang. Robust strategies for mitigating supply chain disruptions. *International Journal of Logistics: Research and Applications*, 9(1):33–45, 2006.

[TG98] T Tolio and SB Gershwin. Throughput estimation in cyclic queueing networks with blocking. *Annals of Operations Research*, 79:207–229, 1998.

[TG09] Barış Tan and Stanley B Gershwin. Analysis of a general markovian two-stage continuous-flow production system with a finite buffer. *International Journal of Production Economics*, 120(2):327–339, 2009.

[TG11] Barış Tan and Stanley B Gershwin. Modelling and analysis of markovian continuous flow systems with a finite buffer. *Annals of Operations Research*, 182(1):5–30, 2011.

[TGYG07] B Tan, SB Gershwin, Rumeli Feneri Yolu, and Stanley B Gershwin. Modeling and analysis of markovian continuous flow production systems with a finite buffer: Methodology and applications. 2007.

[TM01] Valerie Tardif and Lars Maaseidvaag. An adaptive approach to controlling kanban systems. *European Journal of Operational Research*, 132(2):411–424, 2001.

[TN99] K Takahashi and N Nakamura. Reacting jit ordering systems to the unstable changes in demand. *International Journal of Production Research*, 37(10):2293–2313, 1999.

[Tol11] T Tolio. Performance evaluation of two-machines line with multiple up and down states and finite buffer capacity. In *Proceedings of the 8th international conference on stochastic models of manufacturing and service operations*, pages 117–127, 2011.

[Tom06] Brian Tomlin. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science*, 52(5):639–657, 2006.

[TR13] Tullio Tolio and Andrea Ratti. Performance evaluation of two-machine line with generalized thresholds. 2013.

[TW05] Brian Tomlin and Yimin Wang. On the value of mix flexibility and dual sourcing in unreliable newsvendor networks. *Manufacturing & Service Operations Management*, 7(1):37–57, 2005.

[Wer01] Loren M Werner. Analysis and design of closed loop manufacturing systems. Technical report, DTIC Document, 2001.

[WJR90] James P Womack, Daniel T Jones, and Daniel Roos. The machine that changed the world: the story of lean production. *Rawson Associates, New York, NY*, 1990.

[Zip91] Paul H Zipkin. Does manufacturing need a jit revolution? *Harvard Business Review*, 69(1):40–50, 1991.