Analysis of Inspection Policy and Risk in High Product Mix Multi-Stage Flexible Manufacturing Systems Subjected to Sequence Disorder and Multiple Stream Effects

by

Anna Rotondo

DISSERTATION

Presented to the School of Mechanical & Manufacturing Engineering of
Dublin City University
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Supervisors

Dr. Paul Young

Dr. John Geraghty

Dublin City University
School of Mechanical and Manufacturing Engineering

May 2010

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy (PhD) is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No.: 58117695

Date:

Alla Notte

che le cose ci nasconde

e alle Stelle

che la rendono preziosa

“Dum loquimur, fugerit invida

aetas: carpe diem, quam minimum credula postero.”

Quintus Horatius Flaccus

“Live as if you were to die tomorrow.

Learn as if you were to live forever.”

M.K. Gandhi

Acknowledgments

Hundreds of creatures have spangled my transient journey on earth and have made my existence wonderful so far and I wish to thank them all, but I fear I would make my masterpiece unbearably tedious.

I am extremely grateful to Dr. Paul Young for realising my dream of becoming a Doctor. His supervision never faltered and remained focused and inspirational throughout my doctorate at DCU. I greatly appreciate the time he has generously given me in resolving any issues I encountered while under his supervision in a professional capacity carrying out research. I would also like to thank him for showing an interest in my personal well being. Dr. Paul has a very positive outlook on life and this is evident by one of his enlightening statements which I hope he does not object to me quoting, "if you are not happy, you won't make the people around you happy". This statement has become an important part of my life and offers me inspiration on a daily basis. I would also like to thank Dr. John Geraghty for his exceptional guidance during the last three years. Dr. John has been very helpful and has given me an infinite amount of technical support and advice which have been invaluable during the last three years at DCU.

I would like to thank Prof. Eamonn Murphy and Dr. PJ Byrne for the interest they have shown for my work and their invaluable advice on how to improve my thesis.

My transfer from Politecnico di Bari to DCU was made possible by the extreme generosity of Dr. Paul and Dr. John and the openness of my previous supervisors, Dr. Giuseppe Casalino and Prof. Antonio Domenico Ludovico. Both supervisors were truly inspirational in setting out early goals and targets for my academic progression and they have blessed me with fond memories of my years at Politecnico di Bari. Their smiles and those of my ex-fellow students have become impressed in my heart.

I would like to acknowledge and thank the production staff of the company that has supported this research. The level of help and support they offered was tremendous and unfaltering throughout the project. It ensured that my research was at all times very enjoyable and rewarding.

My stay in Dublin over the last few years has been enjoyable and made memorable by a few wonderful people. I'm especially grateful to a girl, who has been my fellow student, neighbour and favourite chef. In her I found my alter ego. I wish her all the best in the years to come.

I wouldn't be on the verge of realising a dream without the unwavering support of my fantastic family. They have inculcated in me the love for life. They have instilled the belief that my dreams should be pursued and should become a reality. I owe a great deal of gratitude to them. The morbid affection we feel for each other sustains me in every moment of my life. The awareness that they are always there for me is a source of immense serenity. Skype has been a faithful companion for the last three years, but in spite of that I've missed them all very much, each day with the same intensity.

Finally, it's not every day that you are the victor having experienced a plumbing catastrophe such as a faulty water pump and an unavailable real estate agent but in this case it was to become a life changing experience. It gave me the opportunity to meet a man whom has since become the love of my life, a man who brings unlimited amounts of joy and happiness into my heart on a daily basis.

Table of Contents

Analysis of Inspection Policy and Risk in High Product Mix Multi-Stage Flexible Manufacturing Systems subjected to Sequence Disorder and Multiple Stream Effects

Anna Rotondo

Abstract

When inspection economies are implemented in complex manufacturing environments, quality risks will arise. The impossibility to predict the monitoring effectiveness of an inspection strategy in all the stations of a production system eventually leads to a loss of time, money and resources which could be avoided. When a product-oriented sampling is implemented in one station of a production segment, the analysis of the available quality measurements presents relevant complexities in all the stations of the segment. The complexities arise as the multiple streams of product and the randomness of cycle times manifest their effects at the stations upstream or downstream of the sampling station. For the sampling station, the variability of the departure process is responsible for the loss of the deterministic pattern of sampling when a global flow perspective is considered.

This research develops fundamental models which support the prediction of the 'quality risk' in all the stations of a high product mix, multi-stage, parallel manufacturing system subjected to multiple stream and sequence disorder effects. The 'quality risk' is measured in terms of number of unsampled items between consecutive samples at a machine level. The time related corresponding measure, that is the time between samples, is partially analysed. Acknowledging the relevance that factory performance decisions have on quality related issues, the impact of some system design parameters on the two performance measures is investigated using a simulation approach. The results obtained have provided a fundamental basis for the development of the prediction models for the distribution of the number of consecutive unsampled items

under different product flow conditions. Based on those models the risk deriving from apparently random sampling can be immediately assessed. The prediction models are also useful in supporting the choice of the sampling parameters able to reduce the 'quality risk' to an acceptable level.

List of Tables

List of Figures

Nomenclature

B	Buffer
IAT	Inter-Arrival Time
L	Large
LS	Line Speed
M	Machine
Pr.	Product
PT	Processing Time
QT	Queueing and transportation Time
Sc	Scenario
S	Small
SI	Sampling Interval
St, Stn	Station
TBF	Time between Failure
TTR	Time to Repair
WIP	Work in Process

Chapter I

Introduction

1.1 Introduction

Quality is an "elusive and indistinct construct" which is often mistakenly reduced to a heterogeneous ensemble of imprecise adjectives [1]. Definitions of quality are often avoided by recourse to uni-dimensional purpose-oriented measures which try to capture circumstantial connotations of a comprehensive concept [2]. Whilst the substance and determinants of quality may elude a systematic definition, its relevancy to companies and customers is unequivocally recognised.

In the manufacturing environment, product's quality is relevantly affected by product design. Recourse to quality function deployment helps engineers in translating customers' needs into product design characteristics. Marketers use conjoint analysis to "*explore the impact of different design decisions on sales, profits and cannibalisation*" so that the optimal set of quality characteristics can be defined. "Design for quality" techniques support the development of products which jointly meet customer requirements and production costs targets [3]. However, product quality is not just a matter of design.

The randomness affecting a production system and the limited reliability of the machines operating in it undermine the stability of the system's performances [4, 5]. As a result, a production process designed to generate products which conform to pre-determined specifications does not always guarantee the desired quality outcome. In order to prevent non-conforming products from being delivered to the next stage of the supply chain, quality inspections are usually performed in a production system. The measurements performed on some quality characteristics of the items produced are not only useful to evaluate the conformance of the items to the designed quality specifications. They can be also used to draw inferences on the quality status of the production process [6].

The presence of several inspection points along a production line contributes to a fundamental improvement of the quality level of production and reduces production waste caused by quality failures of machines involved in the process. However, intensifying quality control inevitably means an increase in quality prevention and appraisal costs and, ultimately, an increase in the cycle time of inspected products. For these reasons, the quality strategy implemented in a production system is often determined as a trade-off between the needs of both quality and production managers [7-9]. In order to allow a reduction of inspection costs and minimise the impact of inspections on cycle times, inspection economies are usually implemented in a production system. These can consist of the reduction of the number of inspection points; as an example, in a serial production segment, just one step can be chosen as a sampling step, which is the step where the sampling decision is made based on an arbitrary sampling scheme. Inspection economies can also be obtained by reducing the frequency with which items are sampled at the inspection points.

Independently of the fashion with which inspection economies are implemented in a production system, they are source of quality risk. Quality risk is here intended as the risk of not regularly having information about the conformance of items to quality specifications and, more importantly, information about the quality status of the machines operating in the system. If monitoring machines eludes regularity, the risk of production waste caused by an undetected failure obviously increases. A possible measure of such a concept of quality risk at a machine level can be the time elapsing between two consecutive pieces of quality information, that is two consecutive sampled items processed at that machine. An alternative measure can be the number of unsampled items consecutively processed at a production machine.

It is understandable that the magnitude of the quality risk associated with a sampling strategy is dependent on the level of complexity of the production system. For a serial production system the effects of inspection economies could prove less dramatic than the impact of a reduced sampling frequency in a job-shop system. This is because of both the level of disorder governing the system, which can be measured with respect to the variation of the item sequence order, and the level of complexity of the correspondence relationships between the machines operating in the system. The first phenomenon is known as sequence disorder effect and it is due to the randomness of cycle times [10]. The second one, which is common in production systems that do not implement deterministic rules for routing items between the machines of consecutive stations, is called multiple stream effect [10]. Both these effects complicate the interpretation of the quality data patterns at the non sampling steps [10].

The primary objective of this research is to contribute to a better understanding of the level of quality risk that inspection economies introduce in complex manufacturing environments. This is achieved through the development of prediction models for two quality risk related performance measures. The quality risk related performance measures analysed here are the number of consecutive unsampled items and the time between consecutive samples at any machine of the system. The prediction models for the quality risk will prove an invaluable decision making support tool for management involved in the definition of sampling strategies capable of guaranteeing to operate under a desired level of quality risk. These models can also be used to facilitate management decisions while assessing the efficacy of the implemented sampling strategy.

The production system under investigation can be classified as a multi-product, multi-stage, parallel manufacturing system; adopting a more recent definition, this system can also be referred to as a multi-product, serial-parallel system [11]. The randomness of cycle time and the lack of deterministic routing decision rules make this production system subjected to the effects of sequence disorder and multiple streams.

A product oriented sampling strategy is implemented in the system. For each product flowing serially through the system, a deterministic sampling interval is set so that for every given number of items of that product consecutively processed by a machine of the sampling station, one is chosen as a sample. The sampling strategy is implemented independently at all the machines of the sampling station.

For such a combination of production system and sampling strategy, the main challenges are

- the identification of the mechanisms which annihilate the deterministic pattern of the sampling strategy at the machines of both the sampling stations and the non-sampling stations;
- the prediction of their effects on the quality risk related performance measure.

The identification of the parameters which mostly affect the performance measures is also a fundamental goal of this research.

There is a strong industrial motivation for the development of this research. It stems from the need of controlling the quality risk in production environments whose dynamics elude a systematic control and for which the cause-effect relationships are difficult to foreshadow. This need is testified by the circumstances why this research was commenced. In a multinational company operating in Ireland, the divergent opinions of production and quality managers about the effects on quality of the variation of the factory line speed highlighted the lack of available models to conduct systematic analysis on the relationship between production system design parameters and quality risk related measures. This contributed to develop the awareness in quality management that the quality risk related performance measures adopted in the firm could be monitored but they could not be predicted. As a result, it was evident that a reactive approach to quality was adopted; the impact of both production and quality related decisions on the quality risk could only be analysed *a posteriori* by using real data. Ultimately, the benefits of having prediction models for the quality risk able to support an *a priori* analysis of the cause-effect relationships were clear.

The academic motivation for this research is based on solid foundations. There are fundamental gaps in the literature relative to two research fields investigated in this work.

Firstly, as highlighted in a recent paper by Jin [11], research focusing on the quality control of serial-parallel multi-stage manufacturing systems is very limited. Despite being very common in reality, serial-parallel multi-stage manufacturing systems are rarely subjects of investigations due to the relevant complexities by which they are characterised. Structural information, such as correlations between stages, and material flow dynamics, such as item sequence disorder, when ignored during the development of quality control strategies impact the effectiveness of the quality control strategy adopted [11]. The analyses available in the literature tend to focus on either the structural information [11-13] or the sequence disorder [14-16]. To the author's knowledge, the only contributions to the quality control analysis in multi-stage serial-parallel systems subjected to sequence disorder and multiple stream effects are by Fan et al. [10, 17-20]. They investigate the quality control problem in complex manufacturing systems by developing robustly designed control charts capable of coping with both the effects.

Secondly, the little attention paid by researchers to the mutual relationship between production system design and quality related issues was highlighted by Inman et al [21]. They demonstrate with different examples from the automotive industry that the interaction between quality and production system design is more important than theretofore recognised in the academic world. Inman's invitation to investigate this relationship in the several different research areas embraced has motivated works in the fields of buffer location [22-26], ergonomics [27], rework policies [28], absenteeism [29], plant build complexity [30], line speed [31]. However, more contributions are needed to fully explore the complex interaction between quality and productivity.

This PhD research contributes to fill the first gap mentioned since it investigates the effectiveness of a quality control strategy in a serial-parallel multi-stage manufacturing system without ignoring the presence of sequence disorder and multiple stream effects.

Relatively to the second gap highlighted, the mutual relationship between quality and line speed is investigated using a different perspective from the one adopted by Inman [21] and Owen [31]. The line speed is intended here as the inverse of cycle time rather than processing time and the concept of quality is expanded so to include control aspects. This confirms Inman's premise that the intersection between quality and system production design can go beyond the domain defined in his paper.

The most interesting elements of this research consist of the availability of a simulation model completely built on real data coming from the company which supported this research and the novelty of the prediction models for the distribution of one of the quality risk performance measures analysed here. In particular, recourse to enumerative techniques to develop distributions resulting from the combination of degenerative distributions represents an interesting approach which could find applications to various problems categories.

1.2 Organisation of this thesis

This section provides a brief summary of the contents of the different chapters which constitute this thesis.

Chapter II presents a review of the literature focusing on the different research issues investigated in this thesis. Analyses on inspection economies from various perspectives are reported. Contributions to the emerging field of the intersection between quality and production system design are explored. Finally, studies about issues related to the flow of material in complex manufacturing environments are analysed.

After a general introduction about the merits and limits of the simulation approaches, Chapter III gives a detailed description of the simulation model developed to investigate the behaviour of a segment of a real production system, from a quality control perspective, under operating conditions which would have been difficult to implement without relevant consequences on the production performance.

Chapter IV introduces the experimental plan used to explore the efficacy of the sampling strategy. Based on the simulation results, the impact of production system design parameters such as the line speed and the station configuration on the two quality risk related performance measures is investigated. The effect of the variation of the sampling frequency on the quality risk is also analysed.

A study on the responsiveness of the sampling strategy to quality failures concludes Chapter IV. This study focuses on the sensitivity of the sampling strategy performances to variations of defect introduction modalities consequent to machines quality break down.

In Chapter V, the simulation results are further analysed to derive prediction models for the average values of the time between samples and the number of consecutive unsampled items. An analytical shape is given to the relationships between the control parameters and the quality risk related performance measures already illustrated in Chapter IV from a partial and qualitative perspective. In order to support the evaluation of the quality risk, a stochastic analysis of the performance measures is needed. This is conducted for the number of consecutive unsampled items under different product flow scenarios. The validity of the prediction model for the non-sampling station is tested against the results obtained by the simulation of a production system with a basic structure. The analysis of different operating conditions and the introduction of input errors in the prediction models will be used to assess the robustness of the algorithms developed to the variations of the hypotheses on which they are based.

The last part of Chapter V is dedicated to considerations on the industrial applicability of the prediction models developed. Using the predicted distributions, the quality risk associated with a sampling strategy can be quantified in terms of maximum number of consecutive unsampled items at a given confidence level. A possible approach to set sampling parameters capable to keep the quality risk in the system under the desired level is illustrated.

Chapter VI presents the discussion of the results obtained in this work. Chapter VII concludes this PhD project and introduces future research directions arising from it.

Chapter II

Literature Review

2.1 Introduction

As stated in the previous chapter, the "quality risk" is the core of this research. The risk of not continuously monitoring the quality status of machines operating in a production system arises from the implementation of inspection economies. The impact that inspections have on time and cost in conjunction with the need of delivering high quality products has historically (See Section 2.2.1) put company managers through a Shakespearian dilemma: to set or not to set quality as a priority? The answer is generally not drastic. Compromise solutions are preferred and trade-offs between quality and cost/production are pursued (See Section 2.2.2). Quality still remains a strategic factor; however, when the attention is re-focused on other issues, the alert on it is relaxed. The point is: "How deleterious is this relaxation in sole terms of quality?" or "What's the impact of economically advantageous sampling on the uncertainty level of the inspection policy?"

In order to answer these questions, in this research, a retrospective approach has been taken against the common trend of including economic considerations when analysing quality related issues. The problem of assessing an inspection policy has been abstracted to a level where quality and risk (See Section 2.2.4) represent the research fulcrum. The abstraction proves fundamental, since when the problem is analysed from its nucleus, the obscure way to the synapses turns into a straight short-cut. This means, considerations on cost or productivity could be eventually made at a stage when the sampling strategy has been already assessed from an impartial and isolated viewpoint. When sampling is implemented in a product oriented deterministic fashion in each machine of a particular station of a production segment, the possibility of predicting the number of items that will be processed between two consecutive samples at any machine in that segment represents a fundamental quality risk related measure. The same could be said for the time elapsing between two consecutive samples at any machine. These measures that are closely related to the quality risk perspective could eventually be included in analyses which look to other specific objectives.

In a flexible manufacturing environment, the complexities arising from the implementation of inspection economies merge with the complexities deriving from the flow of material. When this happens, controlling the quality risk proves prohibitive. The randomness of the cycle time and the combinatorial number of paths which items can follow through a multi-stage, serial-parallel manufacturing system introduce in the system a level of disorder which impacts the hypothetical regularity of deterministic sampling plans. Disentangling the skein made up of item sequence disorder (See Section 2.3.1) and random routing patterns (See Section 2.3.2) is a key through which a clearer vision of the problem can be gained.

The need for a quality oriented analysis which takes hexogen elements into account but keeps its internal focus is actually vivid. When the hexogen elements regard system design parameters, the industrial interest mixes with the academic avant-garde (See Section 2.2.3). Only a few years ago, Inman et al. [21] highlighted the need to investigate the mutual relationship between quality and system design issues. Finding inspiration from the automotive industry, they reported several cases where decisions made about system design affected production quality and vice versa. Noting the lack of literature on those themes, they exhorted researchers to focus their attention on 21 research areas defined by the intersection of quality and system design related issues. Among these areas, the relationship between quality and line speed represents the frame by which the research presented here can be located. The meaning of line speed and quality is here different from what intended by Inman. The interaction between line speed and quality as perceived by Inman regards the realisation that reducing processing times could entail a reduction of production quality for both manual and automatic operations. In this study, the impact of line speed, intended as the speed at which items cross a production segment, on the effectiveness of a sampling policy is analysed. The effect of the configuration of the stations is also taken into account during the analysis. With different meanings conferred to quality and line speed/configuration the horizons of Inman's proposal have been widened. Line speed and system configuration represent two of the research areas individuated by Inman for which the interaction with quality has been analysed here. In order to frame these areas in the wider field of the intersection between quality and production system design issues, contributions directly motivated by Inman's paper to different research areas will also be illustrated in Section 2.2.3.

Section 2.2 offers a review of the literature in the quality control area. Quality is analysed with respect to cost (Section 2.2.2), process design issues (Section 2.2.3) and risk (Section 2.2.4). Section 2.3 gives a general review of the academic solutions to the problem of dealing with the complexities related with the flow of material in flexible manufacturing environment.

2.2 Quality control

When different products compete in the market, quality represents one of the most important factors on which customers base their purchase decision [32, 33]. Independently of the particular customer or the product category, the management of production systems can not ignore this attitude and the decision to make quality a crucial element in the production strategy seems to be the only way to guarantee success. Nowadays, the concept of quality is very broad and assumes different connotations not only to different people but even to the same people at different times [34]. The traditional definition of good quality as "conformance to specifications" has been surpassed by definitions which trigger a more proactive attitude with respect to quality commitment at any organisational level [2].

The development and implementation of different statistical tools and humanistic theories about quality in a manufacturing system will be briefly traced in the next section. The remainder of Section 2.2 will explore the relationships between quality and the main factors that help quality engineers in making decisions about monitoring strategies and quality design. As is evident in the graph reported in Figure 2.1, economical considerations have dominated this decision process in the history of quality control. The histograms in Figure 2.1 are based on data extracted from the web-based discovery platform Engineering Village using as search keywords "optimisation & quality control & manufacturing" for the bars in dark blue and "optimisation & quality cost & manufacturing" for the bars in light blue. A review of the literature on the relationship between quality and costs is presented in Section 2.2.2.

Despite the historical significant predominance of economical considerations during the analysis of quality strategies, a good quality design process should not be confined to its mutual relationships with costs. There is a fundamental need of taking other factors into consideration when quality, and as a consequence quality control, has to be designed [21]. These factors embrace all the different elements involved in production system design (See Section 2.2.3).

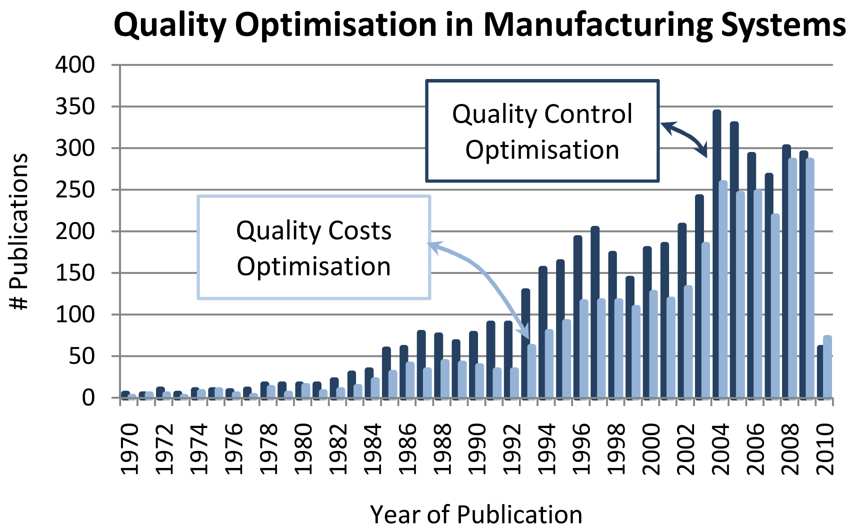


FIGURE 2.1 NUMBER OF PUBLICATIONS ABOUT QUALITY CONTROL OPTIMISATION AND QUALITY COST OPTIMISATION.

Finally, a brief introduction on the concept of risk in quality control will be given in Section 2.2.4. The most common quality related risk measures will be defined. Studies on particular measures that can be easily related to the ultimate objective of this research will be introduced.

2.2.1 Brief history of quality control

The creation of the quality control function in a production system can be traced back to the late 19th century when Taylor introduced the concept of "division of labour" in the industrial world [35]. Along with the standardisation of each process stage, that reduced the single worker responsibility for quality and made management more aware of quality issues. 100% inspection was usually used to guarantee the conformance of the final product to specifications [36].

The use of statistical methodologies for monitoring and improving production quality is largely due to W. E. Shewhart, who in his book "Economic Control of Quality of Manufactured Product" (1931) [37] presented the application of a statistical chart for the control of measurable characteristics of a product. One of the interesting implications of the use of the control charts is the realisation that 100% inspection is not always necessary and the implementation of an efficient sampling strategy can lead to a noticeable cost and time reduction while keeping the system under control.

Different control charts have been developed since Shewhart introduced the first ones. Duncan [38] introduced economic considerations in the design of a control chart; this highlights the ever increasing interest in pursuing quality and financial targets at the same time. The evolution of the control chart characteristics has usually followed the demand of the industry where Statistical Process Control (SPC) was applied. Spanos [39] undertakes a brief summary of the evolution of SPC schemes, and in particular control charts, with a particular focus on their applications to semiconductor industries (Table 2.1).

In the same direction of Duncan's work, is the contribution of H.F. Dodge and H.G. Romig, who soon after the definition of control charts, proposed the use of acceptance sampling in place of 100% inspection [40]. The rationale for promoting the application of acceptance sampling was based on the consideration that a production sample, if reasonably large, homogeneous and randomly drawn, can provide the operators with

TABLE 2. 1 SPC TECHNIQUES ILLUSTRATED BY SPANOS [39].

Traditional SPC techniques	
$\bar{x} - R$ chart	– Controlling location and spread of a continuous variable – Able to detect only large variation of the spread – Suitable for subgroup size less than 10
$\bar{x} - S$ chart	– Controlling location and spread of a continuous variable – Suitable for subgroup size greater than 10
Moving Range chart	– Ideal when data can not be easily grouped – Simple to use – Frequent false alarms due to the possible data autocorrelation
p-chart	– Attribute chart – Controlling the fraction of nonconforming items
c-chart	– Attribute chart – Controlling the number of defects on each inspection unit – Assumption of defects distributed according to a Poisson distribution with a constant defect density
u-chart	– Attribute chart – Controlling the average defect count over a group of n entities
CUSUM chart	– Based on the concept of Maximum Likelihood – Sensitive to small and persistent deviations of a process – Faster response and more unambiguous interpretation than the Shewhart charts
Modern SPC techniques	
Hotelling's T^2 chart	– Based on the concept of Multivariate Control – Sensitive to the collective deviations of a number of cross-correlated parameters from their respective targets – Clear global picture of the process status – Reduced number of false alarms
Regression chart	– A Model-Based SPC technique – Based on the development of prediction model of the parameter to be monitored – Systematic out of controls indicate the need to update the regression models – Ideal for multi-recipe production environment
Time Series Analysis	– Controlling the forecasting errors – Elimination of the problem of the auto-correlation of the parameters measures for continuous parameter readings

enough information to develop inferences about the population from which it was drawn and which it tries to resemble [34, 36]. The debate about the effectiveness of sampling or screening strategies animated research on quality control in the past decades. For multi-stage systems, under particular quality costs scenarios, Lindsay and Bishop [41], White [42], Britney [43] and Raz [44] supported the recourse to 100% inspection rather than sampling; on the other hand, a high emphasis was been placed on the efficacy of sampling plans rather than screening in multi-stage systems [45, 46]. In the literature of late, simulation approaches [47-49] and evolutionary algorithms [47, 50] are used to support the choice of the optimal inspection strategy, among 0%, 100% and sampling, after each production step in multi-stage systems.

Initially, these statistical concepts were not readily accepted. It was only after World War II that the American government almost forced manufacturers to implement the SPC techniques in order to reduce the production of defective items which was systemic during the war. In order to ease the inspection procedures, sampling tables adapted from the one devised by the Bell System, were published, during the war, as a military standard, MIL-STC-105. However, this standard was in practise extended to contracts signed in the non-military industry as well, and the desired quality improvement was soon realised. [36, 51].

More so than the Americans, the Japanese believed in the advantages of implementing SPC tools in the manufacturing environment. The all-embracing role of quality in a production system involving technical, financial and strategic choices, along with the indications on personnel training procedures, is the ultimate message of the "Total Quality Management" (TQM) methods which were proposed by W.E. Deming. With the human involvement suggested by these motivational theories, quality is implemented at all organisational levels and is pursued in an effective fashion, with a constant focus on cost reduction [36].

During the 1980's, the contributions of G. Taguchi, regarding the Design of Experiment (DOE), and Ishikawa, primarily in respect to Cause and Effect Diagram, have enhanced the application of TQM in manufacturing systems.

During the 1990's, the attention for quality control and management was heavily focused on quality certification. This guarantees that a certified firm operates at standard quality levels, from the perspective of both the manufacturers and the customers. In order to promote quality awareness, recognize quality and business achievements of organizations and make the organizations' successful performance strategies known, different awards have been established. The most popular are the Malcolm Baldridge National Quality Award in the US, the Deming Prize in Japan and the European Quality Award in the EU [34].

More recently, there is an increasing attention in the literature on the emergence of new quality philosophies, such as Six Sigma and Lean Manufacturing. Six Sigma inherits from TQM the focus on customer satisfaction, different problem solving methodologies and the recognition that all employees are responsible for quality. It expands TQM's focuses to objectives complementary to quality, such as availability, reliability, delivery performances and after-market services. The Six Sigma metric, that is one of the novelties of this new philosophy, is now extensively applied in a more flexible fashion. The constant search for perfection, through the elimination of non-value-adding operations and the reduction of variability at every opportunity, briefly summarises the fundamentals of Lean Manufacturing. Quality management practices in lean production are based on the concept of Zero Quality Control (ZQC). A ZQC system includes mistake-proofing, source inspection and 100% automated inspection. A fusion of the two philosophies with the consequent creation of Lean, Six Sigma organisations seems to be the latest trend in Quality Management [52].

2.2.2 Quality vs. costs

Looking through the history of quality control a particular research pattern can be noticed: the constant trend to reduce the impact of quality control on time and costs [34]. This is clear in the introduction of the concept of sampling in place of 100% inspection for monitoring the production quality [37, 38]. In fact, apart from plants where quality tests are destructive, the recourse to sampling simply obeys the need for reducing any time delay and cost increase due to the introduction of inspection points at some stages of the material flow in the production system [34]. Another attempt to monitor process quality while aiming for good net profits has been made by Duncan, who has published a long series of studies about economical design of quality control. Before Duncan [38], Girshick and Rubick [53] had studied quality control issues from an economic perspective. Their contribution consists of the introduction of the criterion of the expected cost per unit time in quality control design. Their objective was to optimise quality control from an economical perspective. However, even though extremely valuable from a theoretical point of view, their study has had very few practical applications due to its complexity; that's the reason why Duncan's model has commonly been considered a reference model for economical design of quality control. Duncan's model [38] quantifies the hourly economic loss associated with out of control production in relation to the net profit during in and out of control production and conventional costs, such as inspection costs and investigation costs following a false alarm. The optimisation is clearly focused on the factory profit itself, with very little focus on the consequence that a poor quality production can have on the market.

Over the years, more attention has been given to the customer and the so called opportunity and hidden costs have been taken into account [54, 55]. In fact, the need to satisfy customers represents a very relevant obstacle to the trend of reducing the number and frequency of quality inspections. The optimisation is still mainly based on costs but cost components of a different nature are included in the analysis. Quality costs are evaluated under different perspectives; Chang et al. illustrate how to measure four types of quality costs in multi-stage manufacturing systems [56]. Menipaz proposes a taxonomy of economically based quality control procedures which, among other things, analyses the various costs taken into account [6]. The most common quality cost model is the PAF model, which include prevention, appraisal and failure costs. In this model, the failure costs are further subdivided into internal and external. Variants of the PAF cost model are also available [57].

A considerable number of papers in the literature present the objective to define a sampling strategy able to achieve a trade-off between inspection costs and costs related to the impact of the sampling strategy on the quality level of the production [58-61]. Hsu [62] illustrates a hybrid sampling strategy for a multistage production process which optimises costs by means of dynamic programming. Penn and Raviv [9] solve two quality control optimisation problems. They consist of the minimisation of the expected operational costs under a given production rate and the maximisation of the expected profit where both the quality control configuration and the production rate are to be simultaneously determined. Inman's suggestion [21] that quality and process design affect each other and should be complementary developed is taken into account here. Using a simulation approach, Alfares includes in the economical optimisation of an inspection strategy considerations about safety; in fact, relief valves in a petrochemical plant are analysed and the risk cost are introduced in the objective function along with the inspection and repair costs [63]. Engi'n adapts Duncan's economic control chart methodology for applications to the weaving industry. The methodology is compared with another optimisation model based on a different interpretation of machine efficiency [64]. Ng and Hui define the economical optimum for the number of learning actions to be taken in place of routine rectifying actions when an out-of-control signal occur [65]. The inspection allocation problem in re-entrant manufacturing systems is analysed by Rau and Cho [66]. They develop a GA approach to maximise the total profit and compare its performance with that of the exact approach, based on enumeration, and a previously developed heuristic.

Many sampling strategies are based on the capacity to exploit the knowledge of the production defectivity in conjunction with cost consideration. This is the case of the studies conducted by Oppermann et al. [8, 67] who develop a quality costs model that eases the choice of the sampling strategy by comparing the actual defect rate with the break even rates determined by the intersection of cost functions associated with different strategies. For multi-stage systems a graphical dynamic programming procedure is proposed to assist the decision process. McIntyre et al. [68], Kuo et al. [69] and Hall et al. [70] develop sampling strategies for the maximisation of the excursion detectability with costs considerations. Lin et al. [71] focus their attention on the impact of the defect capture rate of an inspection technology on excursion costs and show the use of a cost analysis program which can provide an optimised sampling strategy based on excursion costs and lots at risk. Jang et al. [72] include yield learning information in their wafer inspection strategy model. Even though more cost intensive in the early stages of a product life, the resulting dynamic sampling strategy, in comparison with its corresponding static sampling strategy, achieves a higher rate of yield enhancement and a reduction in the average cost per wafer in the long run.

Inspector fallibility or, in general, inspection errors and their impact on inspection costs occupy a big part of quality control literature [73-77]. The dynamic sampling strategy suggested by Sheu et al. [78] is aimed to minimise the costs related with inspection and is based on a probabilistic model. Emphasis is given to the impact of inspection errors on the optimal inspection policy and total costs. Ballou and Pazer [79] investigate the impact of the inspector fallibility on the inspection strategy in terms of both error magnitude and variability and propose a sampling strategy that optimises total costs. They find that fallibility magnitude has more impact on inspection costs and configuration than its variability and that it can not be compensated for by an increase in the number of inspections. Moreover, the Type II error is less relevant than the Type I error in terms of impact on the sampling strategy. Type I and Type II errors are classification errors which can occur during an inspection. Type I error refers to the mistake of classifying a process under statistical control as out of control due to its natural variability. It is closely related to the classification criteria used during the inspection. In most cases, a comparison between the quality measurement and a control interval is operated; however, more complicated criteria can be followed. For instance, when control charts are used as quality tools, the Western Electric Company (WECO) rules can be implemented. These rules take into consideration stricter limits than the common 3σ control limits and pay attention to the quality measure patterns in order to detect in advance eventual quality deteriorations of the process. Control charts which adopt WECO rules are characterised by an enhanced sensitivity to quality issues with respect to the charts exclusively based on 3σ control limits. The main drawback of the use of WECO rules consists of the generation of more frequent false alarms, which is extremely inconvenient in terms of time and resources needed for restoring the production. Type II error occurs when an item processed during an out of control scenario is classified as a good quality item because its characteristics are accidentally within the control limits. Lee and Unnikrishnan [80] develop and compare a mathematical model and three heuristic solution methods for the optimisation of sampling plans based on costs and inspection error with inspection capacity and time constraints for a multi-stage, multi-product manufacturing system. Bendavid and Herer try to find an optimal inspection/disposition policy which minimises the expected quality costs associated with classification errors of uninspected items. In order to overcome the limited applicability of the dynamic programming solution proposed, due to its computational complexity, they also develop four heuristics and compare their respective performances [81]. Wang incorporates the effects of Type I and Type II errors in an off-line quality control strategy optimisation model for the minimisation of expected total cost for a batch production [82].

Other topics taken into account for the optimisation of a sampling strategy involve availability constraints and congestion problems at any point in the production system. Sakurai, Fujii and Kahiara [83] base their optimisation on the balance between yield and tool availability. Lee et al. [84] propose a dynamic sampling strategy which would enable the optimisation of the inspection station utilisation by tuning the sampling frequency and the cycle time of inspections according to the length of the queue upstream the inspection. Chen et al. [85] propose a near optimal allocation inspection model for a multi-stage production system with limited capacity and congestion problems for the inspection station. A hybrid sequencing policy is considered ideal for a faster detection of yield problems and congestion costs reductions.

The use of simulation models and several optimisation techniques is obtaining an ever increasing interest. Heredia-Langner et al. [50] uses genetic algorithms and a desirability function to solve the sampling strategy optimisation problem in a multi-stage manufacturing system. The combination of the search and expansion mechanisms used in the solution of the problem is found to affect the results. Van Volsem et al. [47] propose a fusion between a discrete event simulation and an Evolutionary Algorithm in order to model a multi-stage manufacturing system, calculate the costs associated with a sampling strategy and optimise the sampling parameters in terms of inspection location, type and inspection limits. The objective is to minimise the total inspection costs for a given expected proportion of defective items at each stage. Sarhangian et al. [48] solve the same type of problem by means of simulation modelling and a search algorithm which combine Tabu search, Scatter Search and Neural Networks. Vaghefi and Sarhangian revise the model in [48] by including the effect of misclassification errors;

they also conduct a sensitivity analysis to investigate the impact of the distribution of the proportion of defectives and the magnitude of Type I and II errors on the optimal inspection plan [49].

Chan and Spedding [7] have recourse to the combination of simulation modelling, a Neural Network Metamodel, Design of Experiments and Response Surface Methodology to analyse the propagation of defectives in the system. Their model can be used as a decision support tool for optimising the process control configuration of the manufacturing system in terms of quality and productivity at the lowest costs.

2.2.3 Quality and production system design

In direct contrast to conventional wisdom, which held that a product's quality depended on its design more than its production, Inman et al. [21] highlighted the feeling that production system design issues could impact quality in a more significant manner than theretofore recognised in the literature. In fact, only a few studies mentioned the eventual advantages and the opportunities deriving from considering quality and production system design all together [86], even from a performance evaluation perspective [87]. While investigating the impact of Flexible Manufacturing Systems (FMS) on productivity and quality, Chen and Adam [87] conducted first a separate analysis and came to the conclusion that a measure quality in terms of productivity and investments would prove more meaningful and comprehensive. Their suggestion is based on the consideration that an increase in output, usually guaranteed by FMS, has no economic value if it results in a reduced production quality. More generally, Mapes et al. suggest that the correlation between different performances in a company is not necessarily negative; trade-off strategies do not characterise the most successful companies present in their survey [88].

Inman et al. [21] systematically studied the different elements of production system design which interact with the quality performances of a system. That resulted in the identification of twenty-one areas for research opportunity in the field of production systems design for quality and a summary of relevant research in each area. In the intervening period, the intersection between quality and productivity has been recognised as an important research topic as decisions taken in one often impact adversely on the other [28, 86, 89, 90]. The majority of the papers illustrated in the rest of this section explicitly find motivation from Inman's invitation.

Gershwin and Schick [89] provided a taxonomy of quality and quantity issues in manufacturing systems in order to assist researchers entering this emerging research field, which has proved useful in informing the research presented in this project. For two-machine systems in which the first machine is impacted by quality failures and inspection occurs only at the second machine, analytical results of Queuing Network Models (QNM) have shown that in the presence of quality information feedback there are cases when the effective production rate first increases and then decreases with increasing buffer sizes [22]. When the impact of reworking defects produced by stations subject to multiple out of control signals is considered for two-machine systems, it has been shown that there is a buffer level that optimises the effective production rate. Additionally, for these systems it has been demonstrated that improving the failure rate of the first machine does not increase the effective production rate when the buffer level is greater than the optimal one [24]. Similar results were obtained for longer production lines monitored by SPC off-line inspections [25]. A QNM decomposition method for simplifying the analysis of long lines with quality and operational failures by transforming them into long lines with operational failures only has been presented in the literature and comparison with simulations was favourable [23]. In a paper by Carcano and Portioli-Staudacher, the problem of allocating assembly tasks and inspection task in an assembly line is simultaneously solved in a model to balance the line with minimum total costs of quality and installation. The model was tested against two benchmark serial models that pursued the same objectives; the concurrent model always achieves better performances than the two benchmark models [91].

When the impact of production, sales and quality policy development are considered, numerical results indicate that coordinated policies will achieve higher profits than individually deployed or loosely coordinated policies [86, 90]. Inspection frequency and total production run time are simultaneously considered as decision variables in the model developed by Yu and Yu for the profit optimisation of a vendor [92]. Issues arising from the impact of quality decisions on production lot sizing and thereby the performance of the manufacturing system in terms of productivity related performance measures has received some attention in the literature of late. In particular, when process deterioration or machine breakdown are included in the analysis of manufacturing systems, the optimal lot size is not necessarily greater than that obtained from the classical economic manufacturing quantity model [93-95]. Moreover, the presence of preventive maintenance consistently reduces the costs of the manufacturing system [93]. The analysis illustrated in [94] expands the analysis reported in [93] by including the effects of inspection on production lot sizing decisions. Sarker, Jamal and Mondal [96] compare two different rework policies in terms of sensitivity to the production defect level and provide suggestions on which of those is preferable based on production and quality costs considerations. The ratio between the inspection cost and the savings when the inspected item is defective drives the choice of the optimal inspection policy to be applied along with the definition of the optimal production run length in the model developed by Wang [97]. When the restoration cost following a production failure are higher than the defective cost the optimal production run length proves to be longer than the classical length.

Using field experiment studies, Erdinç and Yeow generalise to labour intensive manufacturing processes the cause-effect relationship between ergonomics and quality previously analysed within limited field settings [27] . They find that ergonomics issues facilitate human fallibility and, as a consequence, can lead to a reduction of production quality. They also show how to use quality measurements to draw inference on the efficacy of ergonomics improvement interventions implemented in a plant. Huang and Inman investigate the impact of plant flexibility/build complexity on quality by carrying out a comparative study [98]. Two assembly lines with different levels of task complexity were considered and the performance of the operators in terms of product quality were analysed. At the end of their analysis, they suggest that complexity should be avoided if not rewarded by the marketplace or "*embraced with countermeasures*" whenever positively recognised by the customers. Blumenfeld and Inman demonstrate that quality and throughput suffer from absenteeism by using the results of queuing theory based models [29]. Different models are developed for assembly lines with and without Andon. Random absenteeism proves to be more deleterious than constant absenteeism in terms of both quality and throughput.

The research, just discussed, primarily addresses research opportunities 9, 11 and 14 (rework policies, buffer size & location and production lot size, respectively) identified by Inman et al. [21] on the intersection of productivity and quality. Research opportunity 15 (flexibility) is also addressed in [30]. In research opportunity 4 "Line or Machine Speed", Inman et al. [21] state that "*Further research on how line or machine speed impacts quality would be very valuable to industry*". In their summary of research to-date on this topic they focus on the impact of reducing the time to perform tasks at manual or automated stations on product quality and the impact of task complexity on the trade-offs between throughput and quality performance. On this subject, Boring and Gertman [99] highlight the fundamental role played by the available time to perform a task on human reliability, which obviously affects the delivered quality. Along with the temporal factor, different factors contributing to human error are introduced and their impact on the time needed to correctly perform a task is in turn analysed. In the development of a human reliability model based on the multi-attribute utility analysis, ElMaraghy et al. [100] include time pressure as an attribute contributing to human proneness to errors since it increases the stress level of a task in terms of the available time for its completion. Assuming a relationship between quality and speed based on Taylor's tool life formula, Owen and Blumenfeld analyse the effects of the operating speed on the production throughput performance in the context of processing items in a manufacturing plant. For three different quality policy scenarios, the models developed show that there exists a trade-off between quality and speed which maximise the production throughput. Other quality-speed relationships, not based on Taylor's formula, were explored; similar results were obtained [31].

Although, as stated by their authors, the studies reported in [27] and [29] find inspiration from Inman's paper they do not directly refer to any of the research opportunities identified by Inman. This confirms that the twenty-one research areas described in [21] do not exhaustively describe the variegated aspects of production system design that mutually interact with quality issues.

Inams's attempt to promote a conjoint analysis of quality and production design related issues can be considered as a specific application of the more general considerations by Ackoff [101] on the need for viewing problems "*from as many perspectives as possible"*. Ackoff suggests that, when managing problems in any type of system, a trans-disciplinary approach should be taken in order to avoid to reduce the nature of the problem to the point of view of people operating in the disciplinary category where the problem is initially placed. Indeed, "*disciplinary categories reveal nothing about the nature of the problems placed in them, but they do tell us about the nature of those who place them there*" [101].

2.2.4 Quality control and risk

When quality control is analysed from the quality risk perspective, the most common results recall the well known concepts of customer's risk and manufacturer's risk, used for control chart design. In this case, the quality risk is conceived as the risk determined by misclassification errors which can occur during the inspection process. These errors expose the manufacturer to the risk of incurring in extra failure costs, either internal or external. The Average Run Length (ARL) and the Average Time to Signal (ATS) represent other two quality risk performance measures commonly used in SPC. ARL and ATS measure the expected number of samples and time to observe an out of control signal, respectively, for both the cases when the process is in-control and out-of-control. ARL and ATS can be considered related to the performance measures investigated in this research, which are the number of items between consecutive samples and the time between samples. The idea in common is the quantification of the risk associated with a sampling strategy in terms of both number of items and time. The main difference consists of the perspective they consider. ARL and ATS refer to alarm signals generated by control charts; the measures analysed here simply refer to successive events with a non-negative connotation.

From this point of view, the number of consecutive unsampled items and, above all, the time between samples recall the nature of the variates that are monitored in a newly defined family of control charts. These control charts are based on Time Between Events (TBE), where "time" and "event" can have different interpretations according to the application areas. Chan et al. [102], who introduced this chart in 2000 as the Cumulative Quality Control chart (CQC-chart), refer to the monitored variate as Q and intend it as the number of units required to observe exactly one defect. The flowchart in Figure 2.2 describes the procedure to be followed for the development of a CQC-chart. However, in the case study presented in [102], Q is the observed time between failures, suggesting the versatility of the chart to be applied in reliability studies, as in [103], or in any situation where the random event under investigation can be modelled by a homogeneous Poisson process.

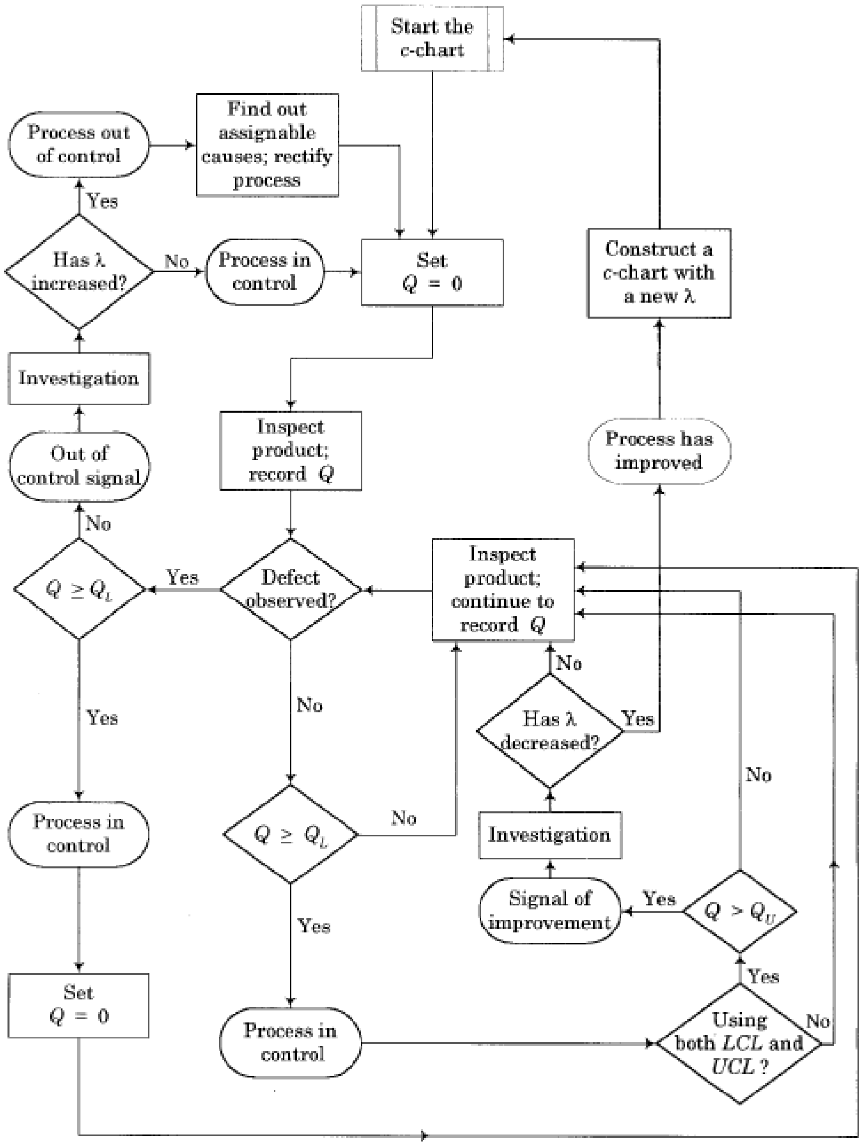


FIGURE 2.2 FLOWCHART FOR THE IMPLEMENTATION OF THE CQC-CHART [102].

The hypothesis of a Poisson process is necessary to guarantee that TBE follows an exponential distribution. The main advantage of this chart is its efficacy in monitoring high-yielded processes, since unlike the c- and u-charts, which are its classical corresponding charts, it does not react to noises so frequently. A standardisation of the CQC-chart has been proposed by Chan et al. [104]; this standardised chart plots the cumulative probability of Q and can be also used for standardising the cumulative count control chart, which unlike the CQC-chart monitors TBE geometrically distributed, generated by Bernoulli processes. Zhang et al. [105] introduce economic considerations in the design of the CQC-chart and define two different approaches for the maximisation of the expected profit per unit time: the pure economic design approach and the economic-statistical design approach, which also considers constraints relative to ATSs. Finally, Shamsuzzan et al. [106] generalise the use of this control chart to multistage manufacturing systems, characterised by multiple streams. They develop a model for the optimisation design of the integrated control chart system, which tries to achieve a proper allocation of Type I error among the individual charts based on the values of the affecting parameters.

2.3 Flow of material

The underlying complexities that govern a multi-stage serial-parallel system represent a fundamental obstacle to a straightforward investigation on the effectiveness of quality control strategies used to monitor the production process [11]. Although very common in the industrial world this type of systems has been object of a limited number of studies [11]. The majority of literature in the field of quality control of complex manufacturing systems dwells on the analysis of inspection allocation and sampling frequency optimisation for multi-stage serial systems [47-50, 107]. Due to productivity and quality balance requirements, several machines are often assigned to a production step and complexities in the control of production quality arise. When parallel machines operate in a serial line, the dynamics relative to the propagation of defects are complex to study. Huang et al. developed a stream of variation model to support the dimensional control in multi-stage serial-parallel systems [12]. They extend the state space modelling approach for a single process route to a serial-parallel system. They also prove that the application of reduction techniques for the elimination of redundant input and output variables make the application of the approach simpler without affecting its capability to capture variation streams. A state-space model is also used by Jin et al. [11] to generalise the chart allocation strategy they developed for multi-stage serial systems [108] to parallel-serial systems. The state-space model is adapted to include the three mechanisms which determine mean shift propagation through the multiple stages of production. These mechanisms consist of the coincidence, divergence and convergence correspondence relationships between machines operating in consecutive stations. The chart allocation strategy developed is applicable to any charting scheme and includes consideration on ARL. The complex interactions among key product characteristics at different stages of a serial-parallel system are analysed by Zeng and Zhou [13]. They use a chain graph building technique that, for each production stage, takes the process physical layout and the relationship found for the previous stages into consideration. Under the assumption that each critical quality characteristic is monitored at each production step, Lam et al. [109] Wu and Shamsuzzaman [110] and Shamsuzzaman et al. [106] optimise an integrated control chart system by allocating the detection power of the control chart system between and within the stages for $\bar{X}$ charts, $\bar{X}$-S charts and time between events charts, respectively.

The analysis of the characteristics of the flow of material in a complex manufacturing system is crucial to the understanding of the mechanism which transforms the properties of a sampling strategy along the different stages of a production line. As a result of inspection economies, in most cases, the measurements taken in a particular production step are supposed to provide information about the quality status of the entire segment. This production step usually coincides with the last step in the segment or, in some cases, it is chosen from among the several steps which constitute the segment for its relative importance in comparison with the others in terms of costs or value added to the processed items. Due to different process flows and routing policies, the cycle time through each segment can vary from item to item; resulting in a disorder effect in the sequence of items between the different process steps. As a consequence, the sequence of measurements obtained in the sampling step might contain misleading information about the quality status of steps upstream or downstream of the sampling station if the dynamics of the flow of material are ignored. Along with the sequence disorder effect, another phenomenon related to the evolution of the flow is relevant in multi-stage serial-parallel manufacturing systems; namely the multiple stream effect. Both these effects shall be illustrated in the following sections.

2.3.1 Sequence disorder effect

For most manufacturing systems, randomness seems to be the word which best describes time related data [111]. Processing times vary not only when different products are manufactured at a particular machine. Variability is intrinsic to most processes and it is usually considered an element to fight in order to ease the control of production and meet quality specifications [4, 5]. For highly automated processes, processing time variability is quite low, and deterministic values do not constitute a poor approximation if a model of the system is developed [112]. Nonetheless, randomness is still present in the system. No matter which queuing discipline is adopted, queuing times are very prone to be variable. In fact, they are highly influenced by machine and resource availability which, in turn, are describable in terms of a random temporal variable. Finally, the item inter-arrival time at each step of the segment is still characterised by randomness [111].

The reduction of variability at any level of production represents a fundamental target in any manufacturing system and not only for quality management reasons [113]. In fact, the Six Sigma objective of reducing the process variance so that the most of the items produced be within the six standard deviations of the mean of product specifications, has proved to provide benefits in terms of processing time variability reduction and cycle time reduction as well [114]. The corruptive impact of variability rather than mean values on system performance is documented for both the processing times [5] and machine availability [4]. The literature concerning the efforts to measure and reduce variability in a manufacturing system is very wide and goes beyond the scope of this study. However, it is worth noting that among the negative impacts that variability has in a manufacturing system there is one which relevantly affects the problem of assessing the risk associated with a sampling strategy; namely the sequence disorder effect [18].

The sequence disorder effect simply represents the variation in the sequence with which the processed items move out from consecutive stations of a production segment. Its impact on the efficacy of a sampling strategy is not banal. In fact, when only one step is chosen as a sampling station, some difficulties may arise in the early detection of out of control production in steps different from the sampling ones, since the sequence disorder effect represents an obstacle to the clear identification of a negative trend pattern because of the data order change. In order to quantify the sequence disorder effect, its magnitude can be calculated as follows. Assume that $\bar{X}$ is the quality characteristic monitored at the sampling station and $\{\bar{X}_i\}$ the item data sequence at the sampling station, with *i* denoting the item output sequence index. When the same data are referred to any other step, *s*, in the production line, the data have to be reordered in order to represent the actual data sequence at that step. If $\{\bar{Z}_k\}$ represents the reordered data sequence with *k* being the item sequence label at step *s*, the sequence disorder magnitude of an item, at the step *s,* can be defined as:

$$D_i = i - k \quad (2.1)$$

This concept is exemplified in Figure 2.3.

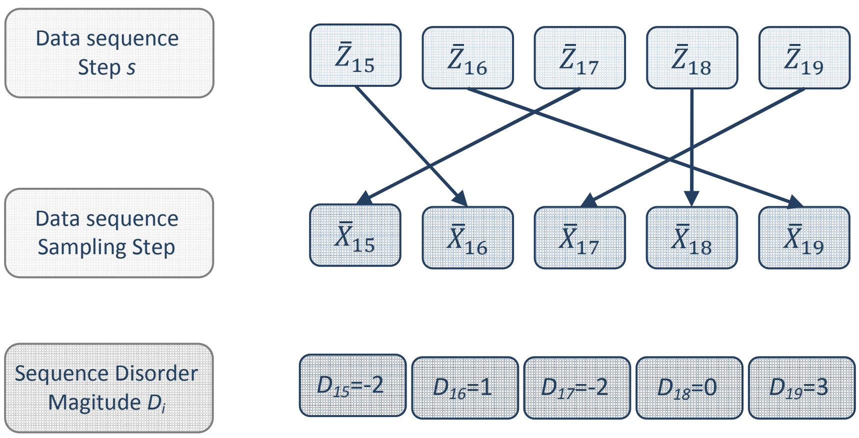


FIGURE 2.3 DEFINITION OF SEQUENCE DISORDER MAGNITUDE.

An overall measure of the sequence disorder magnitude at the step s can be evaluated as the range of the items sequence disorder magnitude:

$$R_s = max_i\{|D_i|\}. \tag{2.2}$$

For the case reported in Figure 2.3 R_s is equal to the absolute value of D_{19}, that is 3; this means that between step s and the sampling step, the system has experienced a variation of item sequence that has involved a maximum of three items.

Only a few studies are available on the impact of the sequence disorder effect on the performances of a quality control plan. The methods suggested to deal with this effect can be summarised as:

- the implementation of data sequence trace back before any statistical analysis [16];
- knowledge-based practices [14];
- the use of several statistics in the control chart system at the sampling steps [10];
- the combination of control charts with variance decomposition as suggested by Montijn-Dorgelo and ter Horst [15] for the distinction of different types of variation in the semiconductor industry;
- the recourse to exponentially weighted moving average (EWMA) control charts with an appropriate weighting factor [17];
- the fusion of classical SPC approaches with data mining techniques, such as predictive analytics [115].

2.3.2 Multiple stream effect

Another complication factor in the analysis of the flow of material in a manufacturing system can be represented by the multiple stream effect. This effect is due to the presence of a different number of machines which can perform the same operation at a specific step. In the absence of any predefined routing policy which would force items to follow a particular path, assigning an item to a machine in the station is usually dictated by machine availability; that is an item, not necessarily the one waiting for the longest time in the buffer, is routed to the first available machine in the station. This usually results in the randomness in the routing patterns; in fact, it often happens that the distribution describing the destination machine from any machine of the immediately previous station is uniform (see Section 3.4.4).

Since the multiple-stream effect introduces more variability in the system, it can be seen as a cause of the sequence disorder [19] and creates remarkable complexity in the phase of tracing back data to the original source of the quality issue. Suppose that a machine m out of the set of identical and independent M machines of a station shifts to an out of control mode. All the items processed by that machine will carry the information about the process shift. However, this information will be spread among the different machines which populate the following stations and it will be difficult to associate the out of control signals with the exact machine which has generated them. For the machines in the stations upstream or downstream of the sampling station, there is not much difference if the sampling is carried out at a machine level or at a station level. However, a sampling strategy implemented at a machine level would avoid the confusing effects of the multiple streams at least for the machines in the sampling station. This last sampling procedure is the most widely adopted in the statistical tools developed for the control of multiple stream processes.

A control procedure for a multiple stream process has a double objective:

- detecting target shifts for one of the streams, assignable causes are affecting only one stream;
- realising when the overall process is out of control, assignable causes interest the overall process.

There are three approaches historically available to monitor multiple steam processes [34]:

- The use of one chart for each stream. The main advantage of this approach is to keep the information about the different streams separate so that a better insight in the process performance is obtained. Moreover, the use of one chart for each stream allows a solution to be derived to the problem of differences in centring, which increases the number of false alarms, and enhance the ability of distinguish between assignable causes affecting one stream or all the streams. The limit of this approach is that for very large systems the number of charts to manage might prove prohibitive to their application;

- The use of a group chart. The information coming from the different streams is plotted on one chart; for the $\bar{X}$-chart the minimum and maximum average for all the streams are plotted with an indication of the stream they are associated with. For the *R*-chart only the maximum value over all the stream ranges is considered. The group chart approach is more convenient than the previous one in case there is a high correlation between the streams; that means the behaviour of the different streams is very similar. However, with this method the insight into the process is lower and the detection capability for target shifts decreases if more than one stream changes at the same time;
- The Mortell and Runger approach, which is a variant of the group chart approach [116]. The main difference is that on the $\bar{X}$- and *R*-charts overall statistics are plotted rather than values relative to each single stream. The insight into the process is even lower and more investigations could be needed if an out of control signal is detected.

The most recent contributions to the multiple stream process control procedure include the development of new statistics to be monitored in place of the traditional ones [117], the integration of more information in the definition of the control limits [118], different sampling strategies [119], economic considerations [106]. Liu et al. [117] introduce four control charts statistics, two based on the F-test and the other two based on the likelihood test, in order to monitor multiple stream processes and distinguish between assignable causes which impact on the overall process and causes which only affect one stream. The main advantage of charts using these statistics is the fact that no historical information is needed. A peculiar positive note of the charts based on the likelihood test is the definition of a specific alternative hypothesis which improves the detection performance of the approach when the observed process changes match the ones which the hypothesis is based on. Contrary to common believe that the use of one chart for each stream could be computationally expensive, Meneces et al. [118] argue that the availability of powerful computer resources makes the running of several charts much easier than the process of decoding and tracing back the information provided by a sample. They suggest that during the definition of the control limits of the different charts, the number of streams and the correlation between them should be taken into account. They analyse the sensitivity .of ARL to the number of stream and the degree of correlation between the streams and find that low correlation reduces ARL when special causes impact all the streams whilst increases ARL when a single stream goes out of control. In the case where a large number of independent streams are present in the system and the lack of full automation introduces difficulties in monitoring each singular stream, Lanning et al. [119] propose an adaptive fractional sampling approach, which consists of monitoring only a fraction of the streams and increasing the sample size, that is the number of sampled streams, only when further information is needed to establish suspected out of control situations. It is noted that an R-chart could join a $\bar{X}$-chart more for improving the detection of differences in the various streams rather than monitoring the variability of the overall process.

2.3.3 Sequence disorder and multiple stream effects

Since, as stated before, the multiple stream effect can be considered one of the causes which generate disorder in the system, it is obvious that the sequence disorder and the multiple stream effects coexist in a complex manufacturing environment. Their combination determines relevant complexities in the understanding of the information provided by quality control procedures.

The problem of dealing with both the sequence-disorder and the multiple-stream effects in a semiconductor manufacturing environment monitored by end-of-line measurements has been investigated by Fan et al. [10, 17-20]. In [17] and [18], the authors propose the use of an EWMA control chart to smooth out the sequence-disorder effect and detect abnormal trends at any process step. Based on the range of disorder between the step to be monitored and the end-of-line, the number of machines in the step and the assumed possible process shift, several possible combinations of the chart parameters are given. The parameters which match the requirements for both the false alarm rate and the quality failure detection speed should be chosen. The main reason for the use of the EWMA relies on the fact that a proper choice of the weighting factor, which is related to the moving average size, allows tuning the relevance of the most recent data. In an environment characterised by a relevant sequence-disorder less importance should be given to recent data, which means it is necessary to operate with a big moving average size, even though a reduction of the detecting speed is obtained. The combined use of EWMA and Shewhart control charts [10] and the fusion of EWMA, exponentially weighted moving C_{pk} (EWMC) and Shewhart control charts [19, 20] are suggested to maximise the detection speed in every condition of process shift. In fact, the EWMA chart outperforms in the case of a small shift ($<1.5\sigma$), the EWMC is the most suitable for the detection of median shifts ($1.5\ \sigma - 2.5\ \sigma$) and the Shewart chart should detect more quickly big process shifts ($>3\ \sigma$). With the SHEWMA (combined Shewhart and EWMA chart) approach, the detection time for small target shifts is reduced by at least 10% in comparison with the combined approach Shewhart-EWMA chart without considering the multiple stream and sequence disorder effects [10]. In turn, the approach considering the EWMC chart as well [19, 20] outperforms the Shewhart-EWMA [10] methodology in terms of detection speed.

It is worth noting that, despite being cited, the problem of the disorder has not been directly addressed by Fan et al. in their papers; rather methods to smooth its presence out have been proposed. Moreover, the end-of-line measurements are taken on a lot-by-lot basis, and the complexity arising from adding the effect of a sampling strategy is not considered.

2.4 Conclusions

The literature reviewed in this chapter focused on the research areas relevant to this research work. The analysis of the contributions to the areas of quality control, with particular attention to its relationship with risk and production system design, and production flow dynamics in complex manufacturing environments highlighted fundamental gaps in research that this study will contribute to fill.

The large number of papers dealing with the optimisation of quality control strategies in simply structured manufacturing systems makes the lack of attention paid by researchers to the analysis of quality control in complex manufacturing systems even more evident. Investigations on the effectiveness of quality control policies in multi-stage serial-parallel manufacturing systems are limited to a few contributions focusing on the development of models able to capture the inter-relation of quality characteristics and the propagation of defects throughout the different production stages. Contributions on the optimisation of integrated control chart systems in multi-stage serial-parallel production environments are also available. Among these, the analyses conducted by Fan et al. [10, 17, 20] provide a fundamental reference for a systematic definition of the flow dynamics that complicate the quality information analysis and contribute to increase the quality risk in multi-stage serial-parallel systems. These dynamics, which can be synthesised in the sequence disorder and the multiple stream effects, originate from the randomness governing complex manufacturing environments. The deleterious effects of randomness determine the lack of control of the regularity of the sampling strategy implemented in a multi-stage serial-parallel system and amplify the risk of not monitoring every single machine operating in it. This risk is the key element investigated here.

The review of the literature on the quality risk (Section 2.2.4) highlighted the great attention paid by researchers to the analysis of risk measures related to the efficacy of the quality control system in signalling suspected quality failures (ARL and ATS). On the contrary, very few studies are available in the literature when the quality risk is measured in terms of the efficacy of the quality control system in monitoring all the processes/machines in the system with the desired regularity. This concept of quality risk, which is adopted in this research, proves particularly interesting since it logically precedes the concept of quality risk on which ARL and ATS are based. It also promotes a more proactive attitude towards quality than the concepts focused on quality failure detection.

The analysis of the quality risk related performance measures illustrated in this thesis will not be confined to quality considerations. It will expand its domain to include the effects of production system design decisions. Only lately, the interaction between quality and production system design has attracted the attention of researchers. Several research opportunities in this hybrid research field have been identified [21] and the number of papers focusing on the mutual relationship between quality and production design issues, as buffer location, rework policies and batch size, is increasing. Ergonomics, workers absenteeism and plant complexity have also been investigated as factors impacting the level of production quality. The negative effects of line speed on the probability of quality defects have been described by Owen et al. [31]. In their analysis, the line speed is intended as the inverse of processing times. The non-exhaustive nature of the investigation proposed in [21] and the fuzzy boundaries of the research opportunities identified encourage to extend the concept of line speed to the inverse of cycle times, so that the effects of queuing time variation can be included. Under these premises, the analysis illustrated in Chapter IV will contribute to the investigation of the mutual impact of quality control and production system design issues from a novel perspective.

Chapter III

System description and modelling

3.1 Introduction

As evidenced in the previous literature review chapter there is a requirement to develop a fundamental understanding of the principle influencers on the risk associated with sampling strategies in complex manufacturing systems. When a sampling strategy has to be defined, the prediction of its performance and the understanding of the impact that some control parameters have on it are highly desirable.

In this chapter, the simulation model used to carry out an analysis of the impact of production systems design and quality control related parameters on two quality risk related performance measures will be described in detail. The performance measures considered support the quantification of the risk of not monitoring the quality status of the machines which populate a production segment and, as a consequence, the status of the items processed by them.

The problem investigated here was inspired by an industrial case. The system initially analysed represents a segment of a wider production line which can be classified as a multi-product, multi-stage, parallel manufacturing system. Based on that segment a simulation model was developed. At a later stage, modifications to the system configuration were also considered in order to analyse the impact on the performance measures of some variables related with the system configuration, such as the number of stations in the segment and the number of machines in a station. In order to abstract the conclusions drawn for the simulated systems to more general scenarios, a basic model, consisting of only two stations and a buffer, was also developed. The simulation results obtained from this basic simulation model proved useful for the investigation on the validity domain of the stochastic model developed to predict the distribution of the number of unsampled items in the non-sampling stations. In particular, the new set of simulation results supported a better understanding of the hypotheses required for the application of the stochastic model developed. For these reasons, the basic simulation model and its output results will be presented in chapter V, after the introduction of the stochastic model developed for the non-sampling station case.

Before describing the development of the simulation model used for this analysis, a brief general introduction on modelling issues will be presented. This is primarily intended to highlight the merits and the limits of the simulation approach in comparison with the analytical approach and hence, justify the choice to complete this research with the development of analytical models.

3.2 Modelling

In order to find the solution to a problem, the system which the problem refers to has to be modelled with a detail level sufficient to guarantee the validity of the solution for the analysis purpose. The system includes all the entities and elements that interact together for the accomplishment of an objective [120]. The constraints and the rules which regulate the system should also be integral part of the model if it is aimed to represent reality as good as possible. Assumptions are usually made during the model development and might prove very useful for simplifying the solution procedures or making the solution possible.

Modelling can be used for different purposes [121]. In analysis, it enables the generation of the output corresponding with a system configuration and a given set of inputs. In optimisation it can be used to optimise the objective function. A model and some solution procedures may support the investigation of the system behaviour and help in the comparison of alternative systems.

Two big families of modelling techniques proved particularly useful aids in this study: the analytical approaches and the simulation approaches. The choice of one or the other approach is usually dictated by several factors, among which the nature of the problem, the complexity of both the problem and the system, the availability and power of solvers represent the crucial factors.

The simulation approach, with its merits and limits, will be illustrated in the next section. Some applications on the research field of quality control, that is the area where this approach has proved to be useful in this study will be introduced in section 3.2.1.2. Consideration on the analytical approaches will be presented in Section 5.1.1 where the analytical model will be introduced.

3.2.1 Simulation

As Shannon [122] states, "*simulation is the next best thing to observing a real system in operation since it allows to study the situation even though we are unable to experiment directly with the real system*". The reasons why a system would not allow a direct exploration of its behaviour under certain hypothesis can be different; it may not exist yet or its manipulation might prove too expensive or time consuming. While the complexity of a system might represent a constraint for the possibility to be modelled by means of mathematical methods, such as algebra, calculus or probability theory, there is almost no limitation to the types of system that can be simulated [121]. A proof of that is the broadness of research areas where simulation has been successfully applied; these areas range from manufacturing to ecology and environmental issues, from business to biosciences.

Among the different simulation approaches available, the discrete event simulation approach has been chosen to conduct part of the analysis presented here. With this technique, the evolution of a system over time is modelled so that changes of its state variables are allowed at discrete points in time. At these points, some events occur and, as a consequence, the state of the system may change. The procedure to be followed for developing a simulation model is almost standardised [111, 121, 122]. The different steps will be briefly illustrated since they constitute the methodological structure which will be retraced later on in this chapter, when the development of the simulation model is presented.

The problem definition represents the very first thing to be carefully defined; a comprehensive knowledge of the project scope is crucial since the purpose of the analysis usually has relevant implications on the model building and experimental design phases. After making sure that all the resources needed for the whole simulation process are available, the system should be defined in terms of the elements to be included in the model and the detail level. That includes the assessment of the simplifying assumptions that can be made without reducing the significance of the model for the analysis purposes. Then the conceptual model can be formulated and a preliminary experimental design developed; in this phase, the systems characteristics to be measured should be clearly defined. The collection of the input data, with their eventual conversion into theoretical distributions, constitutes the next step of the simulation process. Once the model is built in a simulation language, its verification and validation should be carried out. Verification is a sort of rigorous debugging that aims to verify if the computer program effectively implements the features of the conceptual model; on the contrary, validation seeks to show that the model developed validly reproduces the behaviour of the real system. Different techniques of verification and validation are available [123] and should be chosen based on the model characteristics. At this point the experimental phase can start and the simulation process can end with the analysis and documentation of the results.

3.2.1.1 Advantages and limits of simulation

The possibility of modelling even very complex systems with a relevant detail level represents the main advantage of simulation. In fact, in many cases, simulation represents the only technique available for solving very complex problems. For these types of problems, the analytical approach might provide the basic equations to model the system, but their complexity, usually caused by the presence of randomness, might constitute an insurmountable obstacle for gaining either numerical or qualitative solutions. Moreover, if the system does not exist yet, simulation can represent a cost effective way to explore its possible future performances; that still holds for existing systems, when the analysts want to assess the impact of new policies and procedures without interfering with the current operations.

If the simulation approach is used, specific constraints and different rules governing the modelled system can be also incorporated in the model, so allowing a realistic reproduction of the system behaviour. There is no need to force the assumptions on which the system is based in order to keep the system modelling and the solution procedure simple or at least feasible.

The fact that the simulation models usually present a stochastic nature enables a proper analysis of a system afflicted by randomness in any of its events or elements [124]. Analytical models are characterised by a deterministic relationship between input and output parameters. If the model is applied to the same set of input parameters, it produces the exact same response, no matter how many times the solution procedure is applied. The intrinsic randomness on which a simulation model is based generates different results even when the input parameters are kept constant.

The availability of a good simulation model can offer a better insight to the system behaviour and help in the identification of the variables most affecting the system performance. That might, in turn, prove useful for the optimisation of the system performances. In fact, the versatility of simulation models and their capacity of generating outputs relative to a wide input domain make them suitable for being integrated with optimisation techniques. From an optimisation viewpoint, a simulation model can be thought of as a function of an unknown form that transforms input parameters into output performance measures [111]. Considering a simulation model as a function has enabled the use of a family of approaches to optimise simulations based on response surfaces and metamodels [125]. A response surface is a numerical representation of the unknown function; on the contrary, a metamodel is an algebraic approximation of the function itself. Once the function to be optimised is obtained, either in numerical or algebraic form, classical optimisation techniques, such as the random search, the stochastic approximation, gradient-based approaches, response surface methodology, etc., can be applied and an optimal solution can be found.

However, these classical approaches to optimisation prove to be technically sophisticated from the user perspective and they often require a considerable amount of computer time [125]. Nowadays, most simulation software has recourse to a metaheuristic approach for simulation optimisation. This approach considers the simulation model as a black box which produces output responses when input parameters are provided. The metaheuristic approach bases the choice of the input parameters on the results obtained by the previous simulations in a sort of evolutionary methodology. Over the classical approaches, the evolutionary methodology presents the advantage to explore larger domains of the solution space with a fewer number of simulation runs. In order to further reduce the time needed to find the optimal solution, metamodels, usually neural networks, can be also developed and integrated in the optimisation structure. April et al. [125] illustrate the advantages of combining simulation and optimisation with an application to a problem of an investment portfolio optimisation.

Among the principal drawbacks that a simulation approach presents, there are the complexity in the model building itself and the difficulty in results interpretation. Moreover, the simulation process can prove very expensive and time consuming and the main risk, sometimes, is not being able to obtain the desired results in the time available. For this reason, less accurate approaches, such as simplified analytical models, might be preferable.

3.2.1.2 *Simulation and quality control*

Simulation has been applied to several research domains. Particular attention will be given here to different applications in the quality control area. These range from the determination of control chart limits [126] to the identification of problems that affect the productivity and quality of the manufactured items [127].

Roy [126] suggests the use of discrete event simulation as a technique is to explore the actual potential of a manufacturing system and better define its targets. In fact, very often, in control design, and specifically in the developing of the control charts which are supposed to monitor the systems performances, the control limits are based on either historical information or management decisions. In both cases, it is not guaranteed that the true capability of the system is exploited at its maximum level. A very conservative attitude, which tends to look at the past performance of the system rather than at the present when its targets are to be defined, can prevent the realisation of opportunities to gain the improvement margins the system already presents in its structure. The development of a detailed simulation model can help in the evaluation of the best performances achievable with a given system configuration and so can help the management to set objectives at any strategic level. Neubauer [128] uses a simulation approach to compare the performances of the EWMA control chart with respect to other quality-control procedures in medical applications. Goel et al. [127] develop a simulation model of a critical process in a supply chain for "*continuous tracking of product and process quality, cost and time during manufacturing*". The model, which also includes the simulation of the different operator experience levels in detecting quality problems during the manufacturing process, allows the assessment of each process scenario from the perspective of quality, cost and time and it results a fundamental tool for optimising the supply-chain logistics and meet customers' requirements at the lowest possible costs and time. Flowers and Cole [129] use computer simulation to assess the efficacy of different sampling strategies in terms of inspector productivity and average outgoing quality. The implementation of the strategy suggested by the simulation results analysis led to improvements in costs and quality which even though not as good as the ones predicted by the model were still relevant.

Following the trend already found for the SPC tools, simulation applications to quality control progressively incorporate economic considerations. A conspicuous number of publications analyse the costs of quality by means of computer simulations [57, 130, 131]. Freeman [57] exhorts firms' management to use computer simulation methods as support tools for making quality-related decisions and illustrate two different approaches which can be used for simulating quality costs. De Ruyter et al. [132] investigate the impact of inspections and control errors on the total quality costs for an automotive stamping plant monitored by self-inspection system. The model provides an optimal control strategy in terms of number of defective panels to be accepted before stopping and investigating the production line. The simulation also demonstrates the negative effects of pursuing separately quality and productivity targets. In fact, it is found that quality costs are minimised at low line efficiency and, in turn, rapid gains in efficiency can be obtained at the expense of quality. This shows the need of addressing the manufacturing objectives as a system rather than single targets relative to discrete operational areas. Visawan and Tannock [133] try to quantify the benefits generated by quality improvement in the automotive market. They develop simulation models which include manufacturing operations and marketing position with the intent to analyse quality costs and benefits both when the selling price of the products delivered is sensitive to the quality level and when it is fixed. The variable price scenario proved to be better when the organisation operates at excellent quality level; in case of a low quality production, the fixed price scenario guarantees higher profits. Lee et al. [134] propose a cost of quality measurement and estimate it using simulation. The measure helps in the evaluation of the impact of inspection and rework on the quality costs. The model is also used to set quality targets through the implementation of a variance reduction method.

3.3 System modelling

Freely following the steps suggested by Shannon [122], this section will trace the development of the simulation model used for the first part of the analysis. For various reasons, the presentation of some steps might be omitted and the order followed might be subjected to changes. In order to frame the problem investigated here in its environment, the system description, presented in Section 3.3.1 will anticipate the problem definition, reported in Section 3.3.2. The project planning phase is not formally described. When the research proposal was formulated it was evident that the necessary software was already available within the academic institution. The production staff of the company supporting this research, which was familiar with the system modelled, ensured a supportive collaboration during the model development and validation. They also agreed to concede a limited access to the historical database in order to extract the input data needed for developing the model, provided that the information made available would be protected by confidentiality. The conceptual model formulation and the preliminary experimental design are not expressively described in this section; they can be deduced from the simulation model development in Section 3.5 and the experimental analysis reported in the Chapter IV, respectively. The analysis of the real data extracted from the historical database of the company supporting the research is extensively reported in Section 3.4. The model characteristics, the assumptions on which it is based and the complexities in it included are illustrated in Section 3.5. The model validation is presented in Section 3.6. The experimental design and the analysis of the simulation results are reported in Chapter IV.

3.3.1 Description of the system

As stated before, the production line of interest to this research can be considered as a segment of a wider production system. The modularity and repetitiveness with which the system can be described allows restricting the analysis focus on only one segment of the system. Here modularity is intended as the divisibility of the system into smaller parts, or segments, similarly configured and characterised by similar elements. These parts basically consist of production and inspection stations; the transportation system and the buffers between the stations are other elements present in each module.

A production station can perform a particular type of process at different stages of the production. An item is allowed visiting the same station more than once in its production cycle, being processed each time with different operations from the set of operations that the machines in the station are capable of performing. The operation to be performed on an item is usually chosen based on both the production stage, that the item has reached, and its product type.

An inspection station measures some quality characteristics of the parts produced at the upstream production stations. The information coming from that station allows monitoring of the quality status of the whole segment, from both a machine and an item perspective. The fact that the machines in a station can perform different operations means that a station can be shared by different segments. Even if a few segments overlap in correspondence with a station, the focus of one of those segments does not affect the generality of the problem here analysed. In fact, from a theoretical point of view, the segments can be still considered independent of each other and flow is serial between them. The independence between the segments is also confirmed by the way the inspection information is exploited. An inspection performed in a segment provides information about the quality status of the particular operation performed in that "theoretical" segment. Even if the information coming from inspections performed in other overlapping segments was available and useful to determine the quality status of a particular machine of the station in common between the segments, this would constitute an advantage from a control viewpoint. It is worth noting that, even in this case, the analysis performed here, focused on only one segment, still provides interesting results since if some useful information is ignored the quality risk estimates will prove conservative, that is higher than their actual value. However, for completeness sake, the effects of combining inspection data coming from different sources were also analysed and will be shown later on in the course of this thesis (Section 5.4.2).

Different types of products are produced in the system. Relative to a segment, only a few products, at a particular production stage, visit all its stations in a serial order. These products enter the first station, visit all the consecutive stations and then exit the system downstream of the last production station, after an eventual visit to the inspection station. For these products the system layout can be considered a serial production segment. Other products cross the segment in one station and follow different paths in the system. They may re-visit the segment at the same or a different production station during their production cycle. In this case, they would be the products which serially flow through one of the segments which overlap the segment under investigations. For the purposes of this analysis, it is convenient to distinguish between products which serially flow through a segment and products which cross the segment at some stations. The difference between the serial flow types and the cross flow types might not correspond to an actual difference in the product types, since as stated before, products of the same type might revisit a particular station in the segment at later stages in the production cycle. Figure 3.1 shows the product flow dynamics in the system. Station 2 in segment X is also used by segment Y for performing either the same or some other operation type. In order to keep Figure 3.1 as clear as possible, only segment Y crosses segment X; that is not the case in the real system, where more than one segment might cross another one in correspondence with the same station or even more than one station. The different products are represented with differently coloured arrows. As it can be noted, the light blue arrow flows serially in both the segments; however, from the segment X perspective, the arrival of that product in station 2 from station A will be classified as a cross flow.

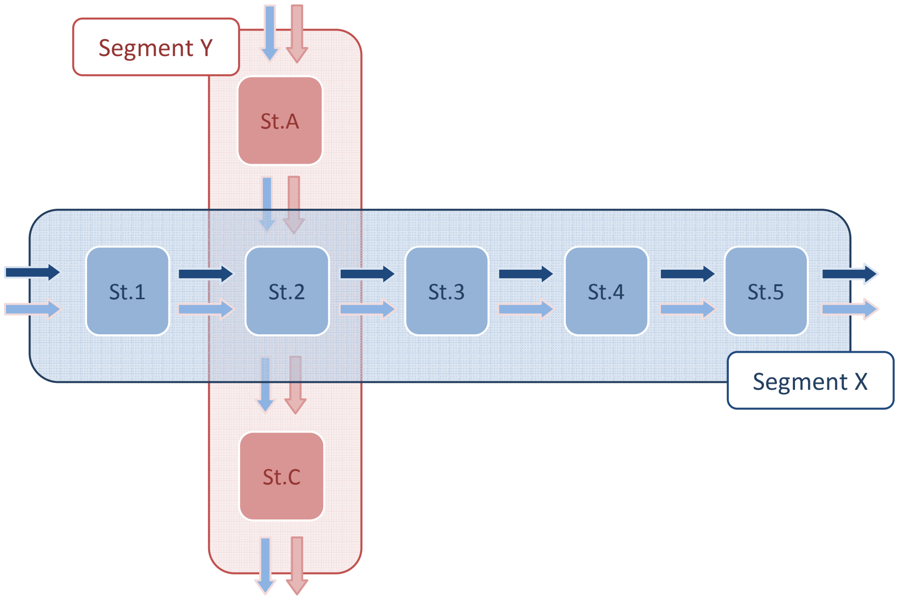


FIGURE 3.1 PRODUCT FLOWS IN THE SYSTEMS. ARROWS OF THE SAME COLOUR REPRESENT THE SAME PRODUCT TYPE.

Each station consists of several machines which operate in parallel. Each of the machines has an independent behaviour and can process more than one item at the same time. The maximum number of items a machine can process in parallel varies depending on the stations. The machines are unreliable and subject to different failure modes. Machines are regularly shut down for preventive maintenance. The frequency for preventive maintenance depends on the stations. Different modes of preventive maintenance are implemented in the system, for example shift-based daily, weekly, etcetera. Each station is provided with an upstream buffer from which machines within the station can select their next items. The assignation of an item waiting in the buffer to one of the machines in the station does not follow any predetermined routing policy; rather it is determined by the machine availability. Transport between stations is via an automated guided vehicle and therefore, transport times are significant.

The configuration of the segment chosen for this analysis is reported in Figure 3.2. The number of stations and the number of machines in each station reflects the operating status of the line when the analysis was initiated. There are two products that flow serially in the full segment; hereinafter, they will be referred to as product A and product B. Each station is also interested by independent cross flows, implying that other products visit the segment. The width of the arrows is approximately an expression of the volumes of the product types represented. Figure 3.2 omits the representation of the inspection station, which for this segment has four parallel machines. It is worth noting that the configuration of the inspection station has a marginal role on this analysis, since, apart from the eventual imposition of inspection capability constraints, which is avoided in this study, the focus is turned towards the production stations. Detailed information on inter-arrival, processing and queuing times will be given in the following paragraphs, along with indications on the availability of the machines.

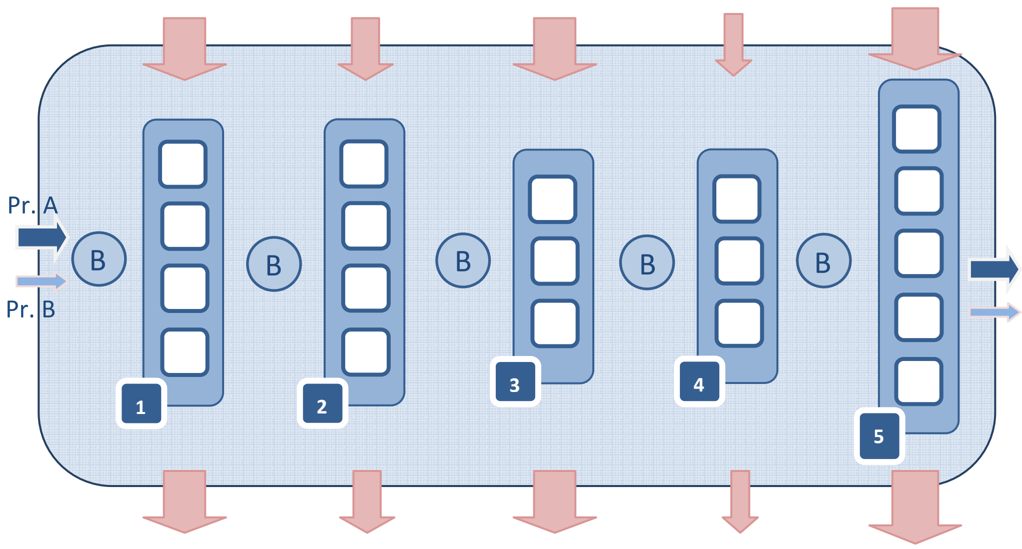


FIGURE 3.2 PHYSICAL CONFIGURATION OF THE MACHINES.

3.3.2 Sampling strategy and problem statement

The inspection strategy implemented in the line is based on a sampling interval for each monitored product. It emulates the skip-lot sampling inspection plan developed by Dodge, which is a sampling plan generally suitable for a continuous stream of lots expected to be good [135]. A particular station, generally the most critical one in the segment, is set as a decision station for making the sampling decision. In that station a particular operation is chosen as the decision point. A sampling interval is determined for each product so that for every given number of items of a given product which sequentially visit a machine in the station for a particular operation one item is marked as to be sampled. The decision is usually made only on the products which follow a serial path in the segment. For reasons different from an ordinary inspection, other products coming from elsewhere in the system visit the inspection station. For the segment investigated here, the sampling station is the fifth station. A sampling interval is set for the two products that flow serially through the segment. The operation chosen is the only one performed on the two products while they flow serially through the segment. That does not exclude the possibility that the two products revisit the same station for different operations; however, in that case they would theoretically belong to different segments.

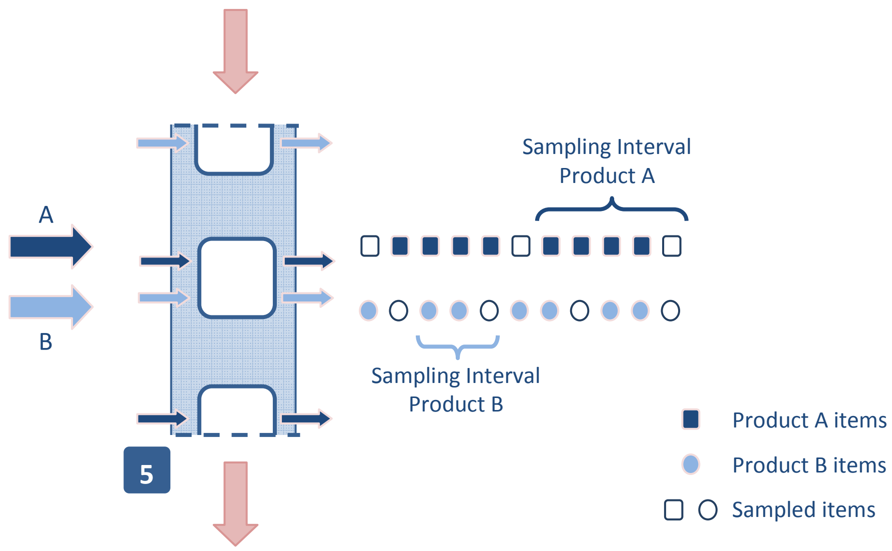


FIGURE 3.3 SAMPLING STRATEGY.

Figure 3.3 depicts the sampling strategy implemented in the segment. The representation of station 5 has been limited to one machine since there is absolute equivalence of the different machines in a station from any viewpoint. The sampling interval of both the product types is shown in Figure 3.3, and it is 5 for product A and 3 for product B. The sampling interval of a monitored product is intended here as the number of consecutive unsampled items plus the sample of that product processed by a machine in the sampling station. This number is deterministic for each monitored product in any machine of the sampling station. However, owing to the randomness of the processing and inter-arrival times, it is easy to foreshadow that once the different product flows in a machine of the sampling station are merged, the count of the unsampled items between two consecutive samples loses its deterministic properties and turns into a random variable whose characteristics constitute one of the fulcra of this research.

The randomness is further accentuated by the eventual presence of a cross flow. In this case, even in the presence of one monitored product type, the number of unsampled items between consecutive samples proves to be random in the sampling station (Figure 3.4).

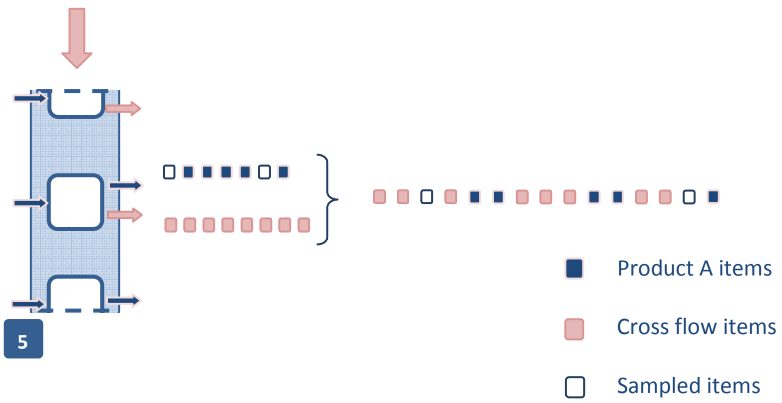


FIGURE 3.4 PRODUCT FLOW MERGING IN THE SAMPLING STATION IN THE CASE OF ONE MONITORED PRODUCT TYPE AND UNMONITORED CROSS FLOW.

As regards the stations upstream or downstream of the sampling station, the so-called non-sampling stations, the situation is clearly less controlled, in the sense that the multiple stream and the sequence disorder effects combine to turn even the simplest case scenario, which would be the one characterised by the presence of only one monitored product type, into a case subject to randomness (Figure 3.5). When the sequence of items processed by a machine of the sampling station is traced back to a machine of an upstream station, it will be evident that the sequence of items at the non-sampling station machine is different; as a consequence, the sampling interval of the items is subjected to possible variations.

The number of unsampled items between consecutive samples has not been the only performance measure analysed in this work. Its corresponding time related measure, the time between consecutive samples has been also considered, at least in the first stages of the analysis. The randomness of this measure is even more relevant since it is a continuous variable and the sampling plan is based on the number of items rather than on the time.

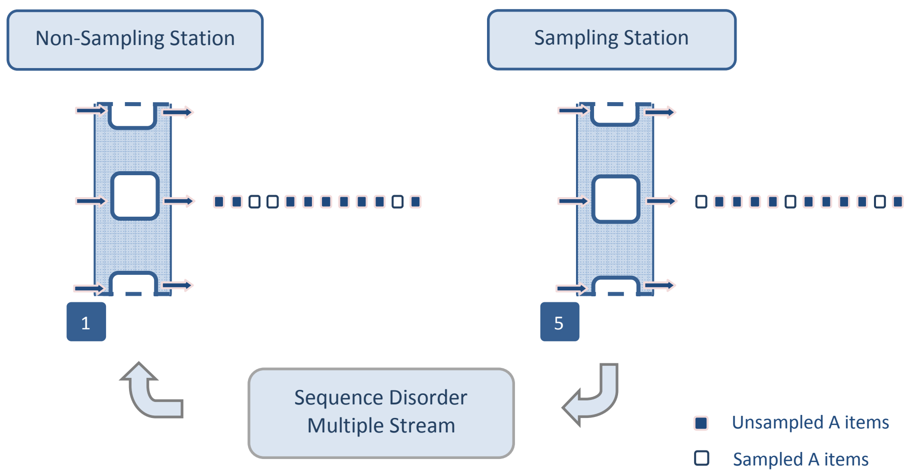


FIGURE 3.5 RANDOMISATION OF THE SAMPLING PLAN IN THE NON-SAMPLING STATIONS.

The aim of this study is to develop fundamental tools for supporting the decision process about the sampling strategy with quality risk considerations. For this research, this ultimately requires the development of some functional models allowing the assessment of the risk associated with a sampling strategy in terms of the aforementioned performance measures. That obviously includes the understanding of the impact of some control parameters on the sampling strategy performances. This last point has been investigated by means of simulation and the results obtained will be illustrated in the next chapter. The control parameters considered are related to the system configuration, the logistic policy, the line speed and the sampling intervals. The prediction models for the distribution of the number of consecutive unsampled items between two consecutive samples will be presented in chapter 5.

3.4 Input data analysis

After choosing the segment to be modelled, the analysis of the data available for that segment constituted a fundamental step in the development of the simulation model. In fact, it was extremely helpful in improving the understanding of the system behaviour and, based on it, discerning which simplifying assumptions were possible and when the model had to faithfully reproduce reality was easier.

The main objective of the data analysis was to model the product flow in the segment as closely as possible; that required taking the variability of the cycle times into account. With this purpose, the different time parameters were analysed not only in terms of mean values but considering their distributions. The next step consisted of fitting theoretical distributions to the empirical data. In order to conduct an objective data analysis, *a priori* assumptions, whether they were justifiable by knowledge-based considerations or by common feelings, were avoided. However, a sort of learning effect was not ignored above all in the cases when the repetitiveness of the nature of the data was not negligible. That reduced the time required to conduct the analysis in the later stages. It is worth noting that sometimes the level of detail of the analysis was conditioned by data availability; in these cases, when even data aggregations didn't yield the amount of data at such a level that the derivation of a probability density function was statistically valid, the time parameter in question was modelled by means of empirical distributions.

The time variables analysed were the inter-arrival times, the processing times, the queuing and transportation times and the availability parameters, which are the time between failures and the time to repair. The next sections will give more details about the procedures followed for the data analysis for each of the time variables considered. The data analysis was mainly carried out using the MATLAB® statistical toolbox, which proved very helpful in selecting the distribution shapes which best fit the data and calculating the corresponding parameters.

3.4.1 Processing times

The Processing Time (*PT*) is intended here as the time spent by an item in a machine in order to receive appropriate operations. It is calculated as the difference between the time at which an item moves out of the machine ($Time_{out}$) and the time at which it moves into the same machine ($Time_{in}$), independently of the number of operations it receives in the same machine

$$PT_i = Time_{out\ i} - Time_{in\ i} \tag{3.1}$$

where the subscript *i* refers to the station which the machine belongs to.

This quantity of time does not necessarily correspond with the actual time needed to complete the operation. In fact, while processing an item, the machine can be subjected to sudden shut down, caused by either preventive maintenance events or unscheduled (random) break-downs. As a consequence, the item waits in the machine for a time much longer than what is actually required to conduct the operation. It is worth noting that, as happens in the real system, in the simulation model these items are kept in the machine in order to wait for the completion of the operation when the machine functionality is restored. The processing times for these particular cases are easily detectable since they generally correspond with the highest values in the list of the processing times relative to a machine. These anomalous values were excluded by the list, since the time delay caused by maintenance events was included in the simulation model in a different fashion.

The entrapment of the items within a machine for maintenance events does not constitute the only reason why anomalous processing times can be found. Extremely low processing times often resulted when incomplete items were processed. This type of item was usually sent to a station in order to assess small variations in the product design. The entries relative to these items were deleted so that the fictitious process time variability caused by them could be ignored.

Other anomalies present in the data could be traced back to the human factor. In fact, it was found that low processing times for standard items were due to the decision of the operators to reallocate the items into a different chamber of the machine where they were originally routed or move them to other machines in the same station. That could be done for different reasons. Similarly, the presence of high processing times might be caused by the delay with which the operator collects the processed item from a machine or by the missing registration of either the $Time_{in}$ or $Time_{out}$ in the database. Errors of an informative nature might be another reason of anomalous processing times. These errors are usually due to different classification criteria used by different operators for the same item.

Finally, the elimination of the anomalous processing times from the dataset on which the fits of the theoretical distributions were based was considered opportune even when these times were not attributable to any of the reasons listed before. Even though those times could have been the expression of the intrinsic variability of the processing times, they appeared excessively far from the typical values and caused relevant bias of the population statistics. The elimination of the anomalous values was carried out separately for each dataset investigated. As a general criterion, the values higher than the 95th percentile of the distribution of the processing times were eliminated; in any case, no more than 10% of the data were excluded from further analysis.

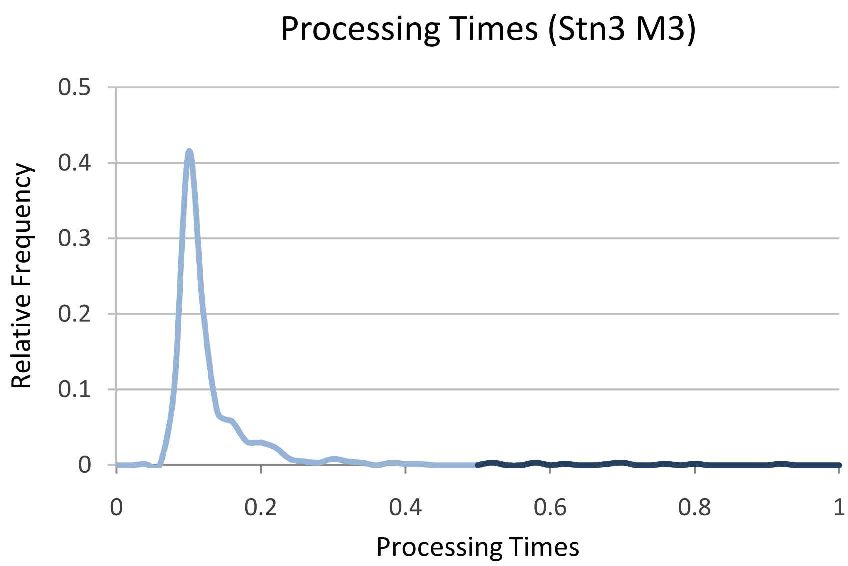


FIGURE 3.6 SAMPLE EMPIRICAL DISTRIBUTION FOR A MACHINE PROCESSING TIME.

Figure 3.6 shows in a darker line the anomalous values eliminated from the list of processing times available for the third machine in the third station. The other part of the excluded values includes the very low values before the peak. The x-axis in Figure 3.6 has been rescaled by an arbitrary factor for confidentiality reasons. The elimination of the outliers in Figure 3.6 causes a 14% reduction of the mean value; the standard deviation is reduced by 50%. It can be noted that the variability is halved whereas the mean slightly decreases. In fact, since values are eliminated from both the extreme parts of the domain the effect of the elimination on the mean can result in either a reduction or an increase; whereas the variability always decreases.

Since no *a priori* assumption was used, the data relative to every single machine in a station were separately analysed. Moreover, for each machine, the data were split based on the operation numbers first and were further divided according to the product. Two categories for the operation numbers were created; the operation performed on the items which flow serially in the segment constitutes one category, the other operations were grouped together since they were not of particular interest for the purposes of this study. Three product categories were considered; they corresponded with the two monitored products which flow serially through the segment and the unmonitored products which cross the single stations. Again, the choice of grouping the unmonitored products is based on the marginal role that these products played in the segment for the purposes of this study. Finally, some categories were grouped; that resulted in the final presence of three macro-categories which are made of:

1. The data relative to the first monitored product (A) for the operation performed while it serially flows through the segment;
2. The data relative to the second monitored product (B) for the operation performed while it serially flows through the segment;
3. All the other data.

This grouping approach did not compromise the modelling accuracy of the processing times and it was considered congruent with the final objective of this study. In fact, all the unmonitored products or the monitored products which do not receive the particular operation performed on the items flowing serially in the segment, from perspective of this study, have the mere role to simulate the reduction of the processing capacity from the monitored items viewpoint. However, at the same time, their presence can not be ignored for the reason that it enables the assessment of the exposition to the quality risk of the unmonitored categories in the segment.

The data analysis performed on the first few machines revealed that the processing times of the first two data categories were significantly similar for the same machine. That was quite expectable since it is reasonable to think that the time needed to complete an operation does not consistently change if the products on which it is performed do not substantially differ from each other and the boundary conditions are kept the same, as happens when a single machine is considered. The statistical equivalence of the two categories as regards the processing time is at the base of the decision to join them for the rest of the processing time analysis. This decision was also partially motivated by the smaller availability of the data belonging to the second category, with comparison with the first one. In fact, in some cases, the amount of available data was objectively limited to make any statistical analysis valid.

Similar processing times for the same operation number were also expected not only in relation to different product categories but with respect to different machines in the same station. The data analysis revealed these similarities; however, it was noted that some groups of machines presented processing times closer between them than times relative to other groups. The differences were not relevant and usually were in respect of the standard deviations rather than mean values. Nonetheless, in order to reflect the variability which characterised the real system in the model, after a machine dedicated analysis, some groups of machines were defined based on the similarities of the analysis results. It was observed that very often the groups of machines based on the processing times similarities found a correspondence with the physical location of the machines. A different position in the plant can mean that different operators with different tasks were looking after the machines; hence, the differences could presumably be traced back to the human factor. However, the operator performances can not be considered the only reason for the difference between the groups since in the time interval considered (six months) it is very likely that personnel turnover was experienced.

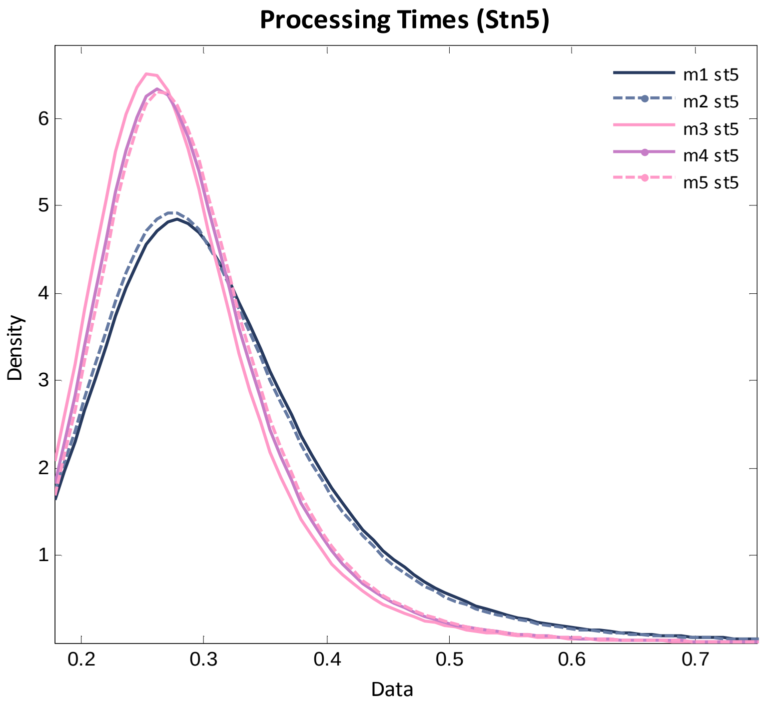


FIGURE 3.7 DIFFERENCES BETWEEN THE PROCESSING TIMES OF TWO GROUPS OF MACHINES IN STN5.

TABLE 3.1 VARIATION OF THE PROCESSING TIME MEANS AND STANDARD DEVIATIONS AFTER THE ELIMINATION OF THE OUTLIERS FOR THE MACHINE IN THE FIFTH STATION.

Machine	Statistics	Before	After
M1	m	0.34	0.32
	std dev	0.39	0.12
M2	m	0.33	0.32
	std dev	0.28	0.12
M3	m	0.29	0.29
	std dev	0.14	0.09
M4	m	0.29	0.29
	std dev	0.11	0.09
M5	m	0.29	0.29
	std dev	0.10	0.08

Figure 3.7 offers a clear example of the similarities of processing times within subgroups of machines in a station. In the case reported, for clarity reasons, the fitted distributions are compared and two different groups emerge. Machines 1 and 2 present a slightly greater mean processing time than machines 3, 4 and 5.

The machine location was not the only possible reason which could justify any difference in the processing times of the machines in station 5. The greater mean and standard deviation for machines 1 and 2 can be also due to the fact that those machine are loaded slightly more heavily than machines 3, 4 and 5 and, hence, the waiting time within the machine itself, which has been always included in the processing time because it is not distinguishable from it, might have increased as a result. Without fitting any distribution, the similarities can be also noted by observing the statistics reported in Table 3.1, before and after the elimination of the anomalous values. Two-way ANOVA has been applied in order to investigate the statistical relevance of both the data elimination and the machine grouping on the variability of the processing times. In order to make the experimental design balanced, the data relative to M4 and M5 have been averaged. This choice is justified by the extreme similarity of the standard deviation for these two machines. The significance levels for the data elimination and the machines groups are 1.24% and 1.33%, respectively. This suggests that both the main effects prove to be statistically significant. As a consequence, the generation of two different groups of machines is a sensible choice. The interaction between the two effects proves less significant (p=3.57%).

Since the grouping based on the different operations and products which generated the third product category was considered general enough, the machine grouping for the generation of a common probability density function for the processing times was not performed for this category. Moreover, the available data quantity for this category was so big that, from a statistical viewpoint, the benefits deriving from a further grouping were not as evident as they were for the other categories. As expected, in most cases, the differences between the fitted distributions were irrelevant, as happens for three machines (M2, M3 and M4) in the second station (Figure 3.8), where the log-logistic distributions are practically indistinguishable; in a few cases, the differences were not negligible, however, they were in respect of the variability more than the mean values

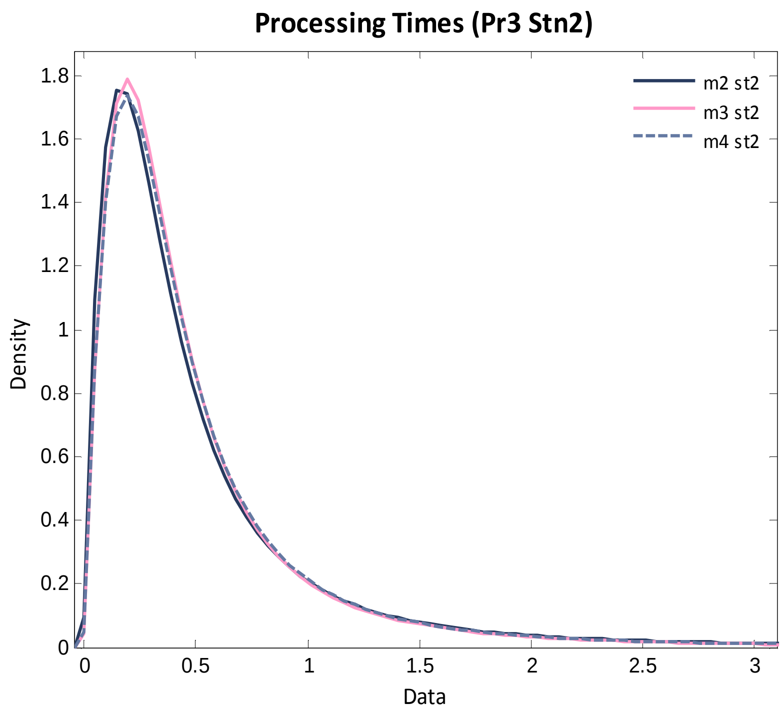


FIGURE 3.8 IRRELEVANT DIFFERENCES BETWEEN THE PROCESSING TIME DISTRIBUTIONS IN STN2.

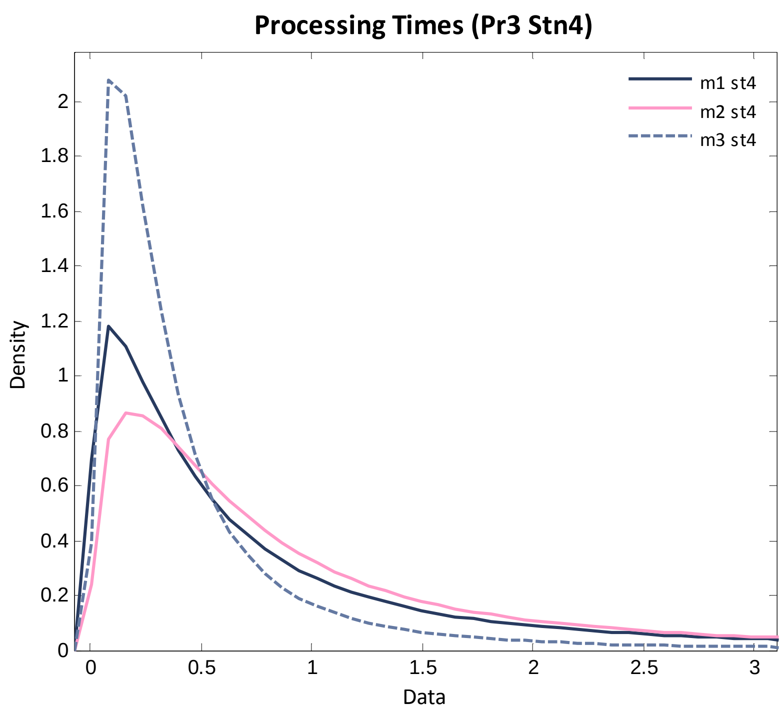


FIGURE 3.9 DIFFERENT VARIABILITY OF THE FITTED PROCESSING TIMES DISTRIBUTIONS (STN4).

(Figure 3.9). For both Figure 3.8 and Figure 3.9 the data are shifted by the minimum values.

The data analysis for the first two categories proceeded with the individuation of the distribution types, with the corresponding parameters, which best fitted the empirical distribution for each group defined. The choice between the different distribution types was based on the log-likelihood ratio.

Data transformations were sometimes needed in order to allow proper data modelling. In particular, data were shifted by their minimum value and the new minimum value, zero, was discarded so that fitting distribution types whose domains were strictly greater than zero (e.g. lognormal, log-logistic distributions) was possible. As is evident, these transformations do not intend to be variance-stabilising transformations. As for the elimination of the anomalous values, these transformations have the objective to make the data smoother and ease the distribution fitting process. This concept applies to the all the time-related input parameter analyses illustrated in this chapter.

Finally, for practicality, one distribution type was chosen for modelling all the data groups. This choice was based on the consideration that the differences between the distribution types were almost irrelevant and that the benefits deriving from dealing with only one distribution shape were significant. The advantages were considered from the perspective of the future use of the model. In fact, a provisional experimental plan included the variation of time related parameters, in terms of both mean values and variability, in order to investigate their impact on the quality risk related performance measures. Modelling data by means of one distribution type simplifies the data management within the model. It eases the comparison between different data groups and enhances the capability of controlling the effects of parameter changes on the distribution characteristics. This last aspect was particularly interesting. In fact, the variation of the characteristic parameters of a distribution might cause different effects on the centrality, dispersion, skewness and kurtosis according to the distribution type. When one distribution type is considered, the parameter variations are likely to determine similar effects on the aforementioned characteristics. The most common best distribution type among the different groups was chosen as the only distribution type to model the processing time. This distribution happened to be the log-logistic distribution. The parameters of the log-logistic distribution were fitted for all the data groups (Appendix A).

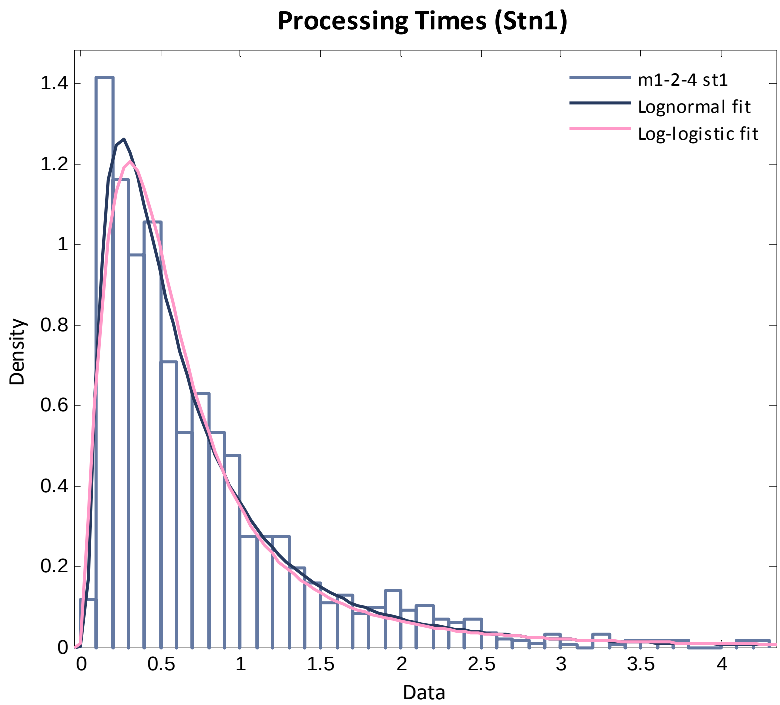


FIGURE 3.10 EMPIRICAL AND FITTED DISTRIBUTION FOR THE PROCESSING TIME RELATIVE TO A GROUP OF MACHINES IN THE FIRST STATION.

Figure 3.10 shows how the lognormal and the log-logistic distributions fit for the empirical processing times for a group of machines in the first station. The data represented were shifted by their minimal value. Based on the log-likelihood ratio, the lognormal normal distribution proved the best one to fit the data; however, as can be noted in Figure 3.10, the fitted log-logistic distribution is not much different from the lognormal distribution, apart from the peak which moves slightly towards higher values.

3.4.2 Queuing and transportation times

The queuing and transportation times were calculated as the time elapsing between two following operations. Using the same notation as the processing times, the queuing and transportation times at the *i*th station, denoted here as QTi, can be expressed as follows

$$QT_i = Time_{in\ i} - Time_{out\ i\text{-}1} \tag{3.2}$$

The choice of joining the queuing time, which is the waiting time in the buffer upstream of a station, and the transportation time, which is the time needed for the item to reach that buffer after moving out the previous station, was somewhat forced by difficulties in gaining access to the information relative to the time when an item entered into a buffer. On the other hand, this choice proved quite advantageous from a modelling perspective since the necessity of modelling the transportation system was avoided. It is worth noting that the transportation system, in the simulation model, would have made sense only for those items which flow serially through the segment. For the cross flow products, modelling the transportation system would have presented more complexities. Nonetheless, for these items the transportation and queuing times have also been calculated in order to simulate the presence of items of different types in the buffer and keep the inventory level under control.

The analysis of the queuing and transportation times was performed as a consequence of the choice of modelling these times as imposed delays on the product flow rather than the consequence of the unavailability of machines ready to process the items in a station.

The approach used in this analysis was similar to the one adopted for the processing times. No *a priori* assumptions were considered for the first station analysed. However, the recurrent results obtained for that station represented a good base on which decisions whether or not to use some simplifying assumptions for the other stations could be easily made.

For the first station the analysis started by splitting the data between the three categories previously defined. The outliers were then individuated and eliminated from the successive analysis. The data elimination almost exclusively regarded high values, since low queuing times were judged not only desirable but possible, given the system characteristics. Since the transportation system was highly reliable, the presence of anomalous values is more likely determined by the human factor impact on the waiting times. In fact, there was no structured queuing discipline implemented in the system, and no pre-determined routing policy either; the choice of the item to be routed to the first available machine in the station was completely left to the operator's discretion. This means that the sequence order of the items moving out of a buffer might be different from the entrance sequence order. An item might be left for a very long time in the buffer and be overtaken by other items more recently located in the buffer. Long waiting times usually represents an anomaly and their inclusion into the dataset can cause relevant bias of the distribution statistics. This is the reason why the elimination of the outliers from the dataset based on which the theoretical distributions were fitted was considered advisable.

Since no priority queuing strategy of any sort was applied in the system, for the queuing times more than for the processing times, the differences between the three categories were expected to be irrelevant. Since the transportation times are included, a higher variability for the third category could be expected for the reason that these items follow different routes. On the contrary, the products of the first two categories visit the stations in the same order and, hence, use the same transport means; as a consequence, eventual differences would be difficult to explain.

When the analysis carried out on the first station confirmed these suppositions, the aggregation of the data belonging to the first two categories was considered opportune. This was also justified by the limited quantity of data available for the second category.

Even if the comparison between the different machines in the first station revealed consistent similarities, it was decided to keep the analysis separated for each machine. In fact, from a statistical point of view, there was no need for further data aggregation.

Since the log-logistic distribution was the distribution that most commonly better modelled the empirical data for the first station, the same distribution type was chosen to fit the queuing and transportation time for the datasets relative to the other machines in the segment (Appendix A). This choice contributes to give a certain level of uniformity to the time parameter modelling and helps for both the data management and the control of the impact of parameter variations. However, relative to the fourth station, for all three data categories, a relevant inadequacy of the log-logistic distribution in modelling the queuing and transportation time was noticed. As a consequence, exclusively for this station, the exponential distribution was preferred (Figure 3.11).

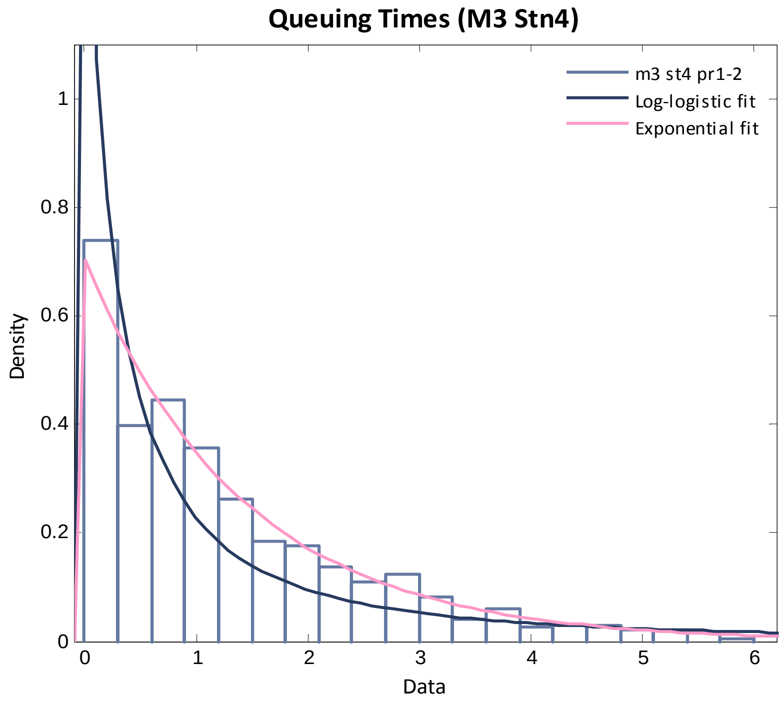


FIGURE 3.11 BETTER FIT OF THE EXPONENTIAL DISTRIBUTION FOR THE QUEUING TIMES IN STN4.

The relative closeness of the waiting and transportation times for the different machines in a station was observed not only for the first station but for all the other stations in the segment. A particular case is represented by the fifth station. In fact, as it was also noticed for the processing times, there are noticeable differences between two groups of machines, above all for the third category (Figure 3.12). The differences can be traced back to the different location of the two groups in the plant; as a consequence, different transportation times are needed to reach each of them. The data reported in Figure 3.12 have been slightly rescaled by an arbitrary factor for confidentiality reasons.

Finally, it is worth noting that tehere are cases when the queuing times are irrelevant in comparison with the transportation times. This is the case of station 3, which happens to share the same working area of station 2. Since there is no transport, the times calculated are reduced to the mere queuing times, which as shown by Figure 3.13,

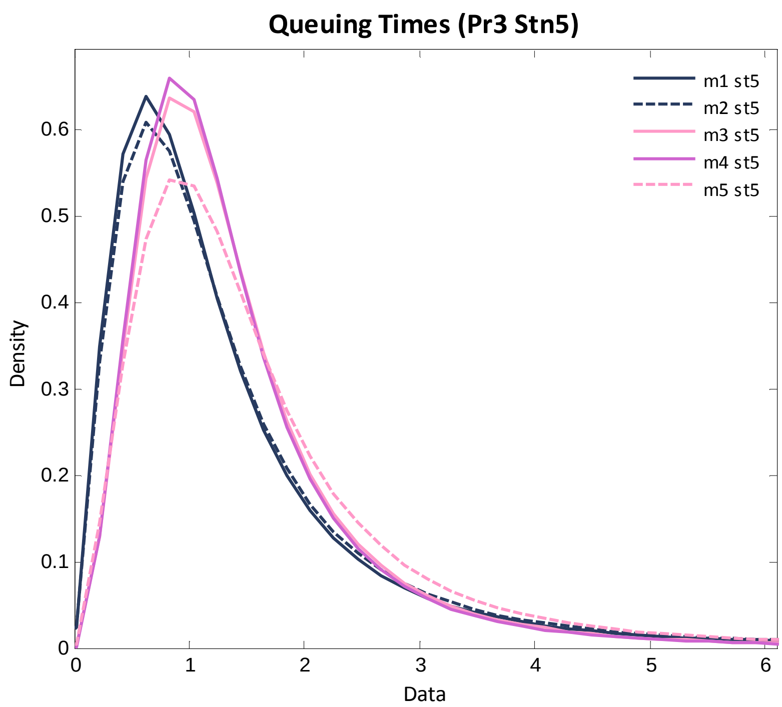


FIGURE 3.12 DIFFERENCES BETWEEN THE QUEUING AND TRANSPORTATION TIMES OF TWO GROUPS OF MACHINES IN STN5.

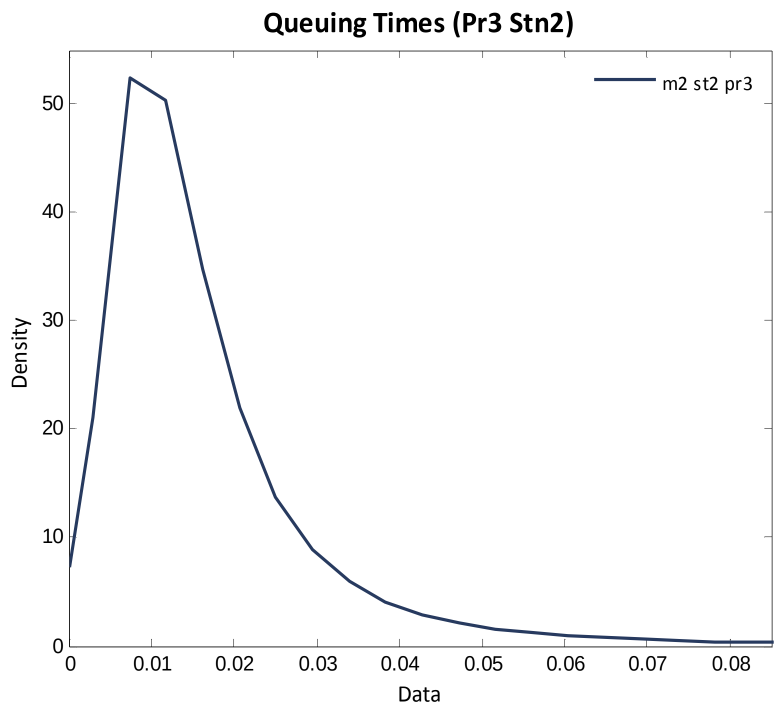


FIGURE 3.13 VERY LOW QUEUING AND TRANSPORTATION TIMES IN STN2.

are extremely low. That happens not only for the first two data categories but for the third one as well. The reason for that relies on the technological link that exists between the two stations so that the products processed in the first one usually visit the second. Since the magnitude of the value was relevant, in this case, the data reported in Figure 3.13 have not been changed.

3.4.3 Inter-arrival times

The inter-arrival time is generally intended as the time elapsed between the arrival of two successive items in the system. Let $Time_{in\ i\ j}$ be the time when the *j*th item in the arrival sequence at the *i*th station moves into that station, then the inter-arrival time associated with that item, $IAT_{i\,j}$, can be formally represented as follow:

$$IAT_{i\,j} = Time_{in\ i\ j} - Time_{in\ i\ j\text{-}1}. \tag{3.3}$$

If all the products flowing through the segment had followed a serial path, the inter-arrival time analysis would have been limited to the first station visited in that segment. In fact, in steady state conditions, the flow in the segment would be exactly steady; that means, the rate at which items arrive at a station should equalise the rate at which they move out of that station. Other conditions would cause unstable situations in the system such as the presence of bottlenecks and the starvation of the machines in some other stations. Moreover, in steady state conditions, the rate at which the items move out of a station should be the rate at which they arrive at the following station, provided that the transportation system is reliable.

However, in the segment on which this analysis is focused, products also exist that cross the segment in a few stations without following a serial route. The fact that those products could visit a few successive stations in that segment has a minor relevance from a modelling perspective; in fact, it might constitute an undesirable and unnecessary complication factor. As a consequence, it was preferred to analyse the inter-arrival times for the cross flows in each station independently of the existence of partial serial routes followed by the cross flow items. On the other hand, the evaluation of the queuing and transportation times for this product category allows the ability, to some extent, to recover the information about the path followed.

With the aim of including queuing and transportation times for all the data categories in the model, the inter-arrival times were calculated as the difference between the times out of the previously visited station for two successively processed items

$$IAT_{i\,j} = Time_{out\ (i-1)^*\,j} - Time_{out\ (i-1)^*\,j-1} \tag{3.4}$$

where *(i-1)** indicates the station immediately previously visited by the item *j*, no matter whether it belongs or not to the segment, and the second sub index, *j*, refers to the sequence of the increasing $Times_{out}$ from the previous stations of the items crossing the *i*th station. In other words, all the cross flow products were grouped and their $Times_{out}$ from the previous station, whichever it was, were sorted in ascending order. Then, the differences between two following $Times_{out}$ represented the inter-arrival times in that particular station for the products of the third category. Appendix A summarises the results obtained by the data fitting process.

With regards to the first station, the first two categories were also considered. Keeping the analysis for the two categories separated is in this case very important, since the difference in the inter-arrival times reflects the difference in the volume fraction of the two product types.

3.4.4 Routing patterns

Even though the routing pattern analysis has apparently little to do with the time-related parameters, the necessity of investigating the presence of preferential paths followed by the items in the segment considered was in part motivated by the approach used for the inter-arrival time analysis. In fact, this was conducted at a station level and not at a machine level. As a consequence, the decision on how to route an item towards a particular machine in a station had to be made. The routing pattern analysis was also aimed to detect any eventual preferential path followed by the items so as to include it in the model. Finally it served to verify the assumption that the machines in a station were uniformly loaded.

For the third product category, the high product volumes made the hypothesis to consider a uniform routing policy between the different machines credible. That means that the items in the simulation model would be routed to a particular machine on a probabilistic basis. In practice, each machine presents the same probability as another in that station to be chosen. However, for the first two categories, it was preferred to conduct an attentive analysis on the routing patterns followed in the segment. The analysis consisted of determining for each machine in a station what the percentages of items moving toward each of the machines in the following station were. The analysis was conducted with the support of Visual Basic macros and interesting results were found.

The most relevant finding that emerged from the analysis was that the results obtained are difficult to generalise. That confirms the assertion by the production staff familiar with the segment that no structured routing policy was applied in the real system. In fact, the decision to route an item towards a machine is left to the operator and it is usually based on machine availability. However, that did not exclude the possibility that some systematic routing decisions could be unconsciously made. In fact, relative to the product categories analysed, some patterns were actually found but, in most cases, there is no apparent reason which can justify them.

The last conclusion is supported by the following results. In contrast to what one may expect, the physical closeness between machines of succeeding stations is not the preponderant criterion followed. Moreover, the fact that the machines of a station belong to the same working area does not generally determine any recurrent routing decision.

This happens, for instance, between the machines of the fourth and the fifth station. Machines number 1 and 2 of the fourth station are located in a different area with respect to machine number 3. For the fifth station, the location of machines 1 and 2 is quite far from the location of the other three machines in that station. It is also far from machine number 3 of the fourth station, which, in turn, is close to machines 3, 4 and 5 of the fifth station. If the location would suggest that the most items from machine 3 of the fourth station would be routed to machines 3, 4 and 5 of the fifth station, the graph in Figure 3.14 shows that product 1 items processed by machine 3 are evenly distributed in two groups which are directed to the two locations of the machines of the fifth station. Then, the items are uniformly distributed between the machines of each location. Since the same considerations can be extended to the other two machines of station 4, it can be concluded that machines 1 and 2 each roughly process 25% of product 1 items coming from the fourth station and machines 3, 4 and 5 process the remaining 50% in equal parts.

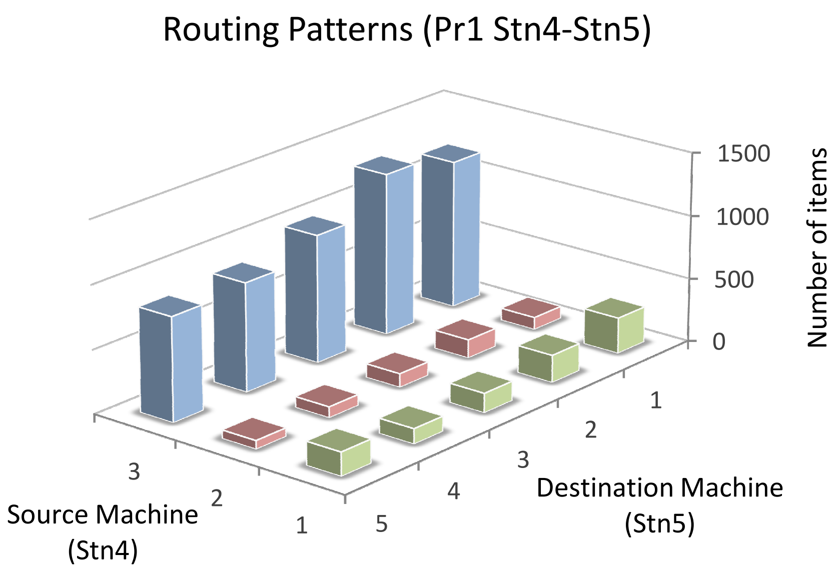


FIGURE 3.14 NUMBER OF PR1 ITEMS MOVING FROM THE MACHINES OF THE STN4 TOWARDS STN5.

Finally, the product has very low impact on the routing rules, as it can be noted in Figure 3.15 and Figure 3.16, which represent the routing patterns between the first and the second station for the items of the first and the second product category respectively. In this case, the items processed in any of the machines of the first station are almost uniformly distributed between the machines of the second station. This situation does not necessarily happen for all the stations. Moreover, looking at the results for the first two product categories, the machines are not equally loaded.

This can cause problems in the evaluation of the inter-arrival times at a machine level; in particular, the linearity between the inter-arrival times at a station level and at a machine level can be undermined if the serial flow is considered. The situation improves when the cross flow is considered. The high volumes which characterise the cross flow and the tendency of the cross flow to spread uniformly across the machines in a station allow the correct use of the linear relationship between the inter-arrival time at a machine level and the inter-arrival time at a station level. The linear factor is given by the number of machines operating in the station.

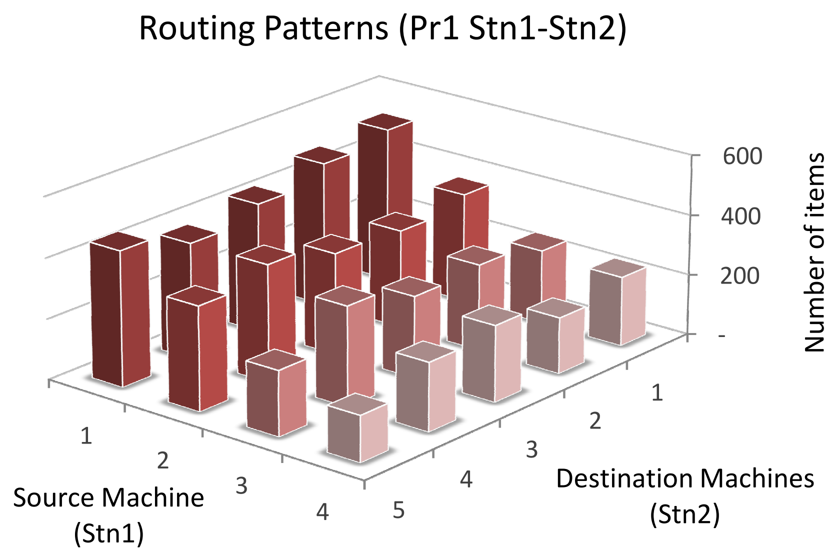


FIGURE 3.15 NUMBER OF PR1 ITEMS MOVING FROM THE MACHINES OF STN1 TO THE MACHINES OF STN2.

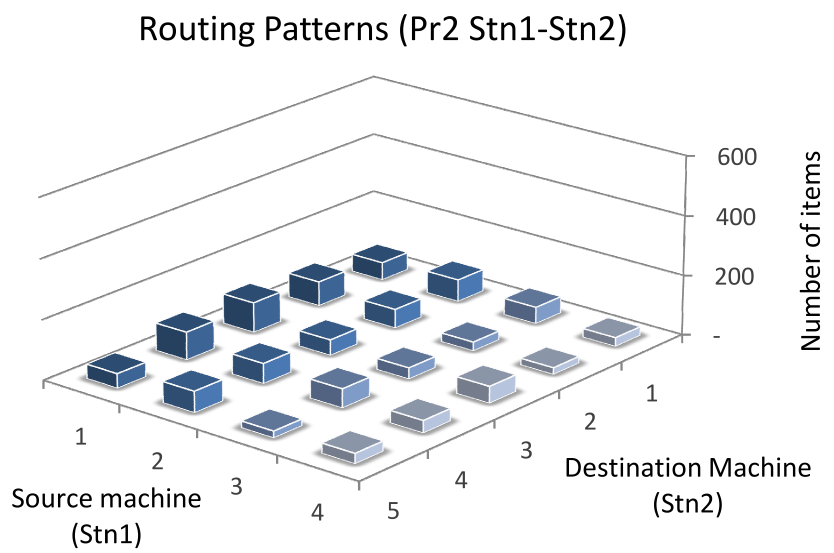


FIGURE 3.16 NUMBER OF PR2 ITEMS MOVING FROM THE MACHINES OF STN1 TO THE MACHINES OF STN2.

It is worth noting that, since the number of items has been scaled using the same factor, Figure 3.15 and Figure 3.16 also show how the first product category is produced in relevantly higher volumes than the second one. In turn, the first product category is noticeably smaller in volume in comparison with the third category.

3.4.5 Availability times

The analysis of the availability times of the machines operating in each station consisted of the analysis of two different parameters, the Time between Failures (TBF) and the Time to Repair (TTR). The TBF describes the time elapsed between two successive shut down events of a machine; the TTR measures the time needed to complete a repair event. A machine can be shut down for either Preventive Maintenance (PM) or sudden breakdown, also known as Corrective Maintenance (CM). The regularity which characterise PM events, in terms of both the availability parameters, is lost when sudden breakdown events are considered.

The availability parameters were analysed at a station level. This was primarily done to overcome the issue of a limited quantity of data available for each machine. It was also justified by the nature of the data analysed. In fact, for the same type of machines which perform the same operations, the PM programs should be the same. Moreover, since the machines of a station share the same working environment and are, pretty much, uniformly loaded, it is very likely that they are interested by breakdown events with similar patterns.

The data available for each station were initially analysed at an aggregated level. This means that no difference was operated between CM events and PM events or between the different PM events. This was due to the difficult interpretation of the information available in the historical database. In fact, it might happen that PM events, while in progress, might reveal the presence of failures of different natures and are turned into CM events. Moreover, the classification of the maintenance events does not follow rigid criteria, so that the same event type can be recorded under different denominations according to the operator which performed it. The distinction between events of different nature was performed at a later stage, limited to the stations for which the aggregated data pattern revealed irregularities which could be presumably solved by a more accurate data classification. This was the case of station 2, where a peak in the upper tail of the distribution of the aggregated data suggested the presence of PM events which could be grouped in a separated data category (Figure 3.17). The time between failures of the PM events were modelled by means of a normal distribution; the exponential distribution fitted the time between failures of the other events very well. It is worth noting that the apparent shift of the fitted distributions towards higher values is caused by the data separation; the time elapsed between two events obviously increases when, in the same time interval, the events pertaining to a data category are reduced.

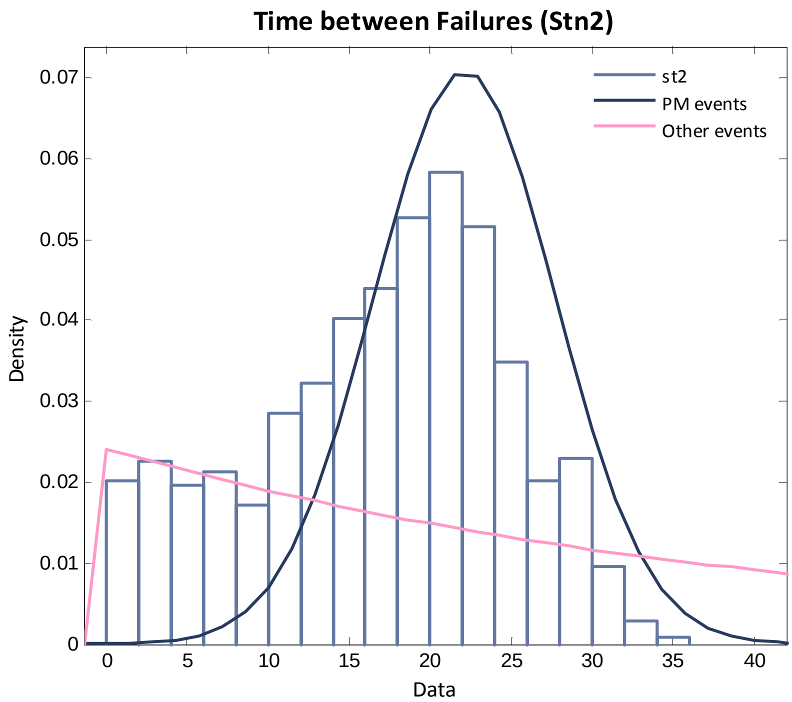


FIGURE 3.17 IRREGULAR SHAPE OF THE TBF DISTRIBUTION AT AN AGGREGATE LEVEL IN STN2 AND THE TWO FITTED DISTRIBUTIONS OF THE CATEGORIES DERIVED.

The exponential distribution, opportunely shifted, proved the best fitted distribution for the data left at an aggregate level. For clarity, in this case, the different events of the maintenance program were not analysed separately and the distributions characterising each of them were not found. As a consequence, there is no evidence that the availability times of each single event were exponentially distributed. This is confirmed by the shape of the availability times at an aggregated level; if these data were the combination of exponentially distributed data, they would follow a gamma distribution.

For the data categories corresponding with PM events the recourse to empirical distributions generally proved the most effective solution to cope with the issue of very low data dispersion.

3.5 Simulation model development

A discrete event simulation model of the production segment introduced in Section 3.3.1 was developed by means of the discrete event simulation software Extend® v6. The graphical interactive approach was used; the blocks available in the different libraries of the software proved to be enough to correctly model the real segment. As a consequence, there was no need to program for developing new blocks or enhancing the functionalities of the existing blocks. In order to keep the model structure simple, the different stations, including the monitoring station, were modelled as hierarchical blocks which were serially located in the workspace so as to resemble the theoretical structure of the segment. Two more hierarchical blocks, one at the beginning and the other one at the end of the line, were also included in the model; they do not represent stations but blocks functional to the items generation and the simulation data management. Figure 3.18 shows the high level structure of the simulation model, with the eight serial hierarchical blocks.

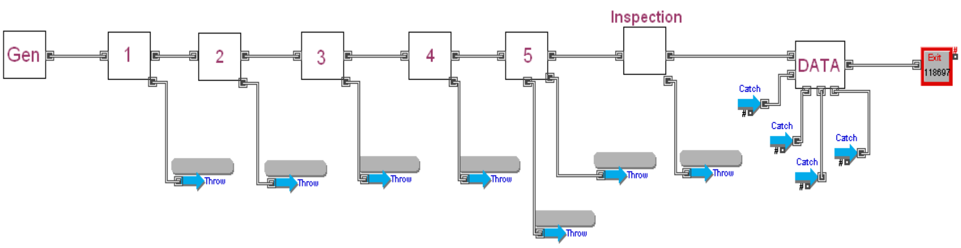


FIGURE 3.18 HIGH LEVEL STRUCTURE OF THE MODEL.

The first block was conceived for generating items which are going to visit the first operation station. It consists of three different generation modules, as can be seen in Figure 3.19. Two of them generate items of the first two product categories. The generation of these items happens only in this first block; once they leave the first operation station they will continue visiting all the other stations serially. The third module simulates the generation of items from the third category which after having been processed in the first station leave the segment and go directly to the last block.

The last block was built with the intention to manage the data collected by the items during their path through the segment. All the items eventually pass through this block. The information gathered is saved in a global array and eventually transferred to Excel worksheets which are generated at the end of each simulation run.

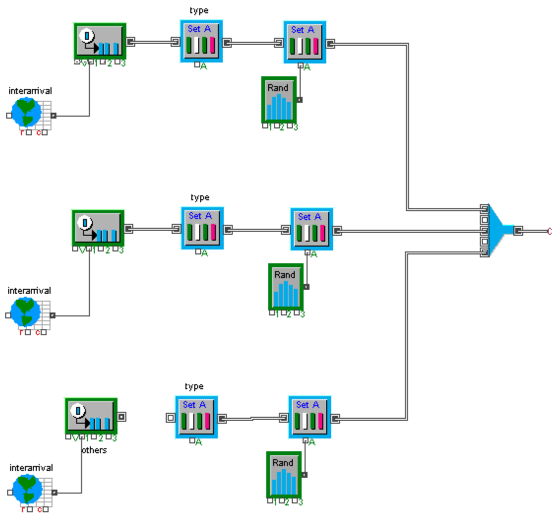


FIGURE 3.19 THE THREE GENERATION MODULES INSIDE THE FIRST HIERARCHICAL BLOCK.

In each generation module the items are generated at random time intervals according to the inter-arrival time distributions fitted on the real data. Soon after the generation, some attributes are assigned to each item. These attributes constitute part of the data which will be recorded in the last block; they include the product type, the timestamp of the time at which the item is generated and the id of the machine that the item is going to visit in the first station. As illustrated in the previous paragraph, this last attribute is randomly generated based on the empirical probability distributions.

Excluding the block representing the first operation station, similar generation modules to the ones present in the first block can be found in all the other blocks representing stations. They only generate third category items, whose path in the segment is limited to the block where they are generated. These items also visit the last block for data collection.

Six hierarchical blocks are located in series to the first block. They represent the five operation stations and the final inspection station. These blocks are similarly structured. After the generation module, a combination module merges the serial and the cross flow items which visit the station. Then, a routing module helps each item to be routed to the machine previously assigned to it. A defined number of machines, each of which is modelled as a hierarchical block, populate the station. A final routing module routes the items to either the following station or the data collection block according to the product type. The structure of the first station is reported in Figure 3.20. The white blocks in the figure represent the machines in the station; an extra machine was included in case the number of operating machines in the station needed to be changed. The routing module described previously prevents the items from being assigned to the non-operating machines.

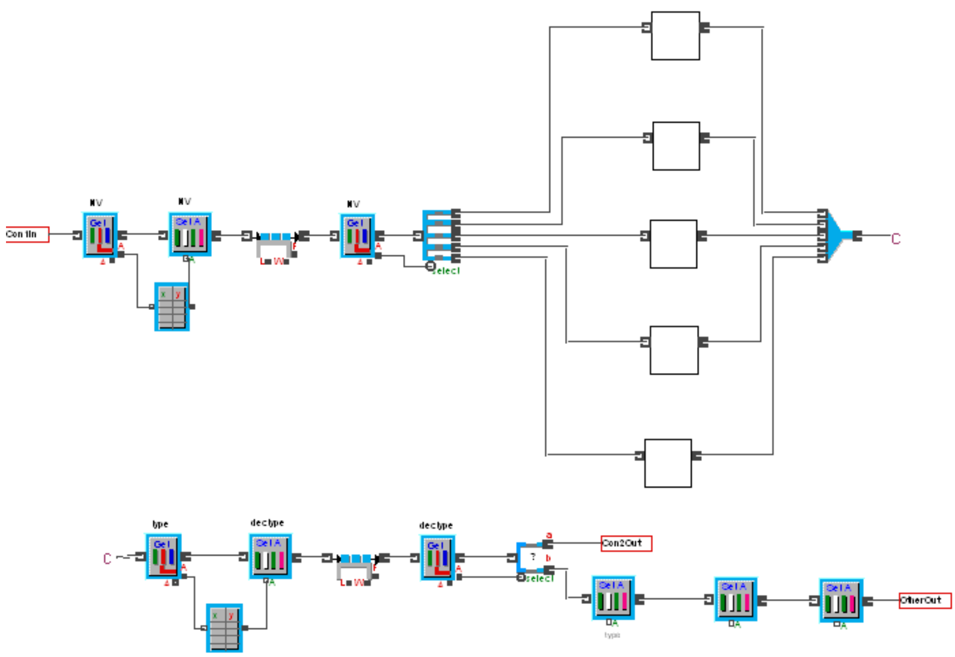


FIGURE 3.20 STRUCTURE OF THE FIRST PRODUCTION STATION.

Each block representing a machine includes two multiple activity blocks. The first multiple activity block simulates the delay caused by the transportation and queuing time. The second multiple activity block introduces the delay caused by the processing time. The imposed delays are provided by a random number generator block.

A shutdown block, located between the two multiple activity blocks, prevents items from accessing the machine when events that compromise the machine availability occur. This block receives signals from different sources, which respectively represent the different maintenance event types identified during the data analysis. Each source sends at a random time frequency a signal which contains the information about the duration of the event. The frequency and the duration of a shutdown event depend on the typology of the event itself.

A timestamp is assigned to each item both when it moves in and moves out from the second multiple activity block. Finally, before leaving the machine, based on both the product and the machine itself, the serial flow items will record the id of the machine they are going to visit in the following station, which is randomly generated based on the routing patterns empirical tables.

The block structure of the machines in the last operation station slightly differs from the other ones because of the presence of a final module intended to implement the sampling strategy. This module is of interest only for the first two product categories. According to the product type, a sampling rate is imposed so that every predetermined number of items of a product type processed by that machine an item is marked to be measured. The marked items will obviously be routed to the monitoring station. All the other items are sent directly to the data collection block.

A more detailed representation of the model structure is available in Appendix B.

3.6 Model validation

The simulation model validation was mainly based on a face validity approach [123]. The production staff familiar with the segment modelled attentively analysed the simulation results and agreed that the simulation model was able to faithfully reproduce the real system behaviour. Attention was first paid to the correctness of the flow dynamics in the system. The availability of graphical animations during the simulation runs supported a better understanding of the logics behind the routing decisions implemented in the model. The flow of items through the segment did not encounter unexpected delays. The simulated availability of the machines operating in the different stations was considered compatible with the maintenance program performed in the real system. This last aspect was of primary concern since it was deemed possible that in the stations where a complicated maintenance program was implemented, blocking issues could arise for the modality with which the combination of CM and PM events was modelled.

The most relevant aspect of the validation process was based on the verification of the typical values for both the performance measures. According to the production staff, the average number of consecutive unsampled items and the time between samples calculated from the simulation results for the initial scenario analysed were reasonably close to the values which they would expect to see in the real system. The model was also able to capture the differences in the monitoring ability of the sampling strategy between the different stations; indeed, the different volumes of items processed at the various stations impacted the magnitude of the quality risk related performance measures in a fashion that the production staff deemed realistic. More importantly, observations concerning the strategy with which the quality risk was monitored and kept under control in the real system most definitely revealed to the production staff involved in this research the suitability of the simulation model for conducting the needed quality risk analyses. These observations involved the quantification of the quality risk in terms of maximum number of consecutive unsampled items. The values of the risk measure calculated for realistic confidence levels by using the distributions derived from the simulation results corresponded to the maximum numbers of consecutive unsampled items which quality management had set as a risk threshold value. When the number of consecutive unsampled items at any machines of the segment reached the threshold an immediate sampling of an item processed at that machine was forced in the real system. This was generally implemented by sending ahead an item to the inspection station. The threshold values adopted were based on a balanced combination of common sense and historical data analysis. The simulation results and the definition adopted for the quality risk could provide a formally correct support to the quality management decisions. For confidentiality reasons, the numerical aspects of this specific validation analysis are not reported.

In general, a validation of the model on a numerical base was made difficult by both the limited access to the company's historical database and some modelling choices. For instance, the choice to model the queuing times as imposed delays prevented the possibility of using the comparison between actual queuing times and queuing times resulting from the simulation as a validation criterion. However, for modelling reasons, a fictitious buffer had to be placed before the block simulating the actual buffer. Hence, an alternative comparison of actual queuing times and simulated queuing times is possible. If the real system is correctly simulated, the time for which items await in the fictitious buffer should prove irrelevant in comparison with the waiting times in the actual buffer; ideally, these times should tend to zero. A preliminary analysis of the queuing times in the fictitious buffers revealed that rarely, usually less than the 5% of the production time, items stopped in these buffers for a time interval. In general, it was noted that a queue built up in the fictitious buffers whenever a machine in the successive station experienced a prolonged shut down. This is understandable since the rigidity of the routing policy implemented in the simulation model. The decision to reproduce the routing patterns found in the real system by pre-assigning to an item the machine at which it had to be processed in the successive station caused a rigidity in the routing decisions which was not experienced in the real system. Whenever a machine is not available, an operator would re-route an item to another available machine in the same station; in the model, re-routing decisions are not possible and an item assigned to a machine temporarily shut down for maintenance operations will wait in the buffer until that machine is made available again. The choice of setting a maximum number of items that a machine can contemporarily process higher than the actual number is partially motivated by the rigidity of the simulated routing policy. Indeed, when the machine is made available the queue built up in the fictitious buffer will reduce in a shorter time and the system will shortly acquire its natural actual behaviour. As a consequence of these modelling choices, a little distortion of the distribution of the total queuing times could be justified.

Figure 3.21 shows the probability plot for the distribution of the global queuing times observed at M1 of Stn4 with respect to the exponential distribution fitting the actual data. The plot has been obtained using the statistical software MINITABv14®. As can be seen in Figure 3.21, the variation of the distribution statistics in the global queuing time distribution is not significant; the mean value of the distribution is reduced by 3.7%. This result suggests that the presence of a fictitious buffer has very little impact on the variation of the resulting distribution of the queuing times since it could increase but never reduce the global waiting times. In this case the variation of the distribution parameter is most likely a mere effect of the randomness of the imposed queuing times. The small value of the Anderson-Darling test statistics, equal to 4.61, generates a small *P-value* for the goodness-of-fit test. The exponential distribution used to model the actual queuing times can be considered a good approximation of the global queuing times derived from the simulation results only when significance levels less than 0.05% are deemed acceptable. Analyses conducted on different machines revealed that the distribution of the resulting queuing times does not significantly differ from the actual data distribution.

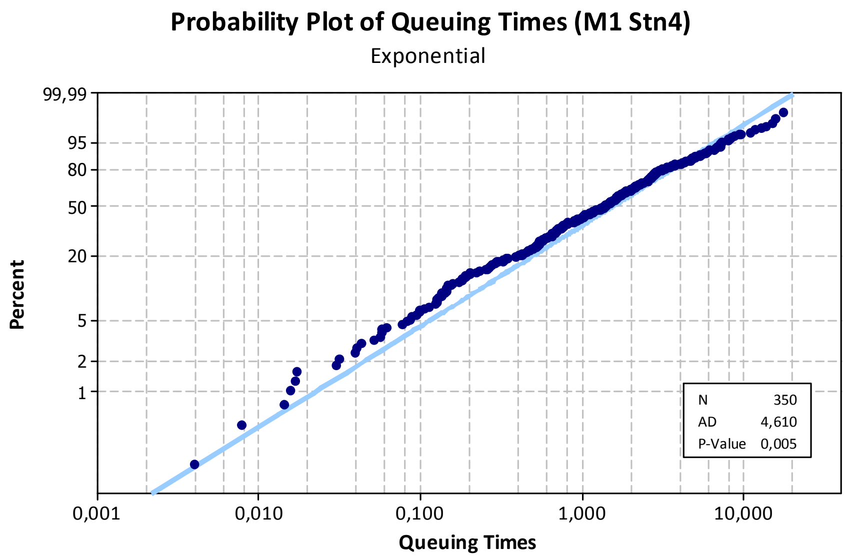


FIGURE 3.21 PROBABILITY PLOT FOR THE DISTRIBUTION OF THE QUEUING TIMES RESULTING FROM SIMULATION AT M1 OF STN4.

3.7 Conclusions

In this chapter, the problem investigated in this thesis was stated. Two quality risk related performance measures under investigation were introduced and the phenomena which determine their randomness, for the particular sampling strategy adopted in the system, were illustrated. The relationship between the two performance measures and some process design parameters and the development of analytical models for allowing quality risk considerations represents the ultimate objective of this work.

The development of the simulation model used for the analysis of the relationship between the two quality risk related performance measures and some process design parameters has been described in details in this chapter. The model reproduces the behaviour of part of a multi-product manufacturing system. The part analysed can be considered as a segment consisting of five operation stations serially located and a final inspection station. Several identical machines operate in parallel in each station. Two types of product flow cross the segment: a serial flow, that interests the whole segment, and a cross flow, that visits only one station of the segment and continues following random paths in the system.

Particular attention has been paid to the data analysis. For the time related parameters used in the model the procedure followed to fit theoretical distributions on the data available in the company database has been illustrated. Data were usually grouped in three macro categories. These were based on the relevancy of the product and the operation types.

In order to simulate the flow of the serial products in a realistic fashion, the routing patterns of the items between the machines of consecutive stations were also analysed. Apart from a few machines for which preferential routes were detectable, the item routing seems to be random; in fact, uniform patterns were often obtained.

The validation of the model mainly relied on the positive feedback of the production staff familiar with the segment modelled. For the limited access to the company's historical database, a validation based on a numerical base proved quite difficult.

Nonetheless, quantitative verifications on the correctness of the model were successfully conducted.

Chapter IV

Quality Risk and Process Design Parameters: a Simulation Approach

4.1 Introduction

The simulation model illustrated in the previous chapter represents an extremely useful tool for investigating the relationship between some control parameters and the quality risk related performance measures considered in this research. In the area of SPC, and in particular of control charts, the ever increasing attention paid to quality risk related performance measures reveals the importance of controlling the quality risk associated with the implementation of a sampling policy. The vastness of the literature reporting analyses on Type I and Type II errors, also called α and β risks, is emblematic in this regard; moreover, the Average Run Length (ARL) or the Average Time to Signal (ATS) are cited very often and are objects of attentive studies [136-138]. In most cases, the analyses take only quality control aspects into account or, at most, economic considerations are included. As suggested by Inman et al. [21], quality issues should never be considered separately from process design elements, and even if the authors refer to quality in general, it is very presumable that their suggestion can be extended to quality control.

In this chapter, an analysis of the impact of quality control and production system design related control parameters on two quality related risk performance measures is carried out using a simulation approach. The performance measures considered allow the risk of not monitoring the quality status of the machines in a production segment to be quantified. As a consequence, the status of the items processed by them can be assessed. This analysis is also preliminary to the development of analytical prediction models for the same measures and it proved very useful in gaining an insight into the system behaviour and determining the parameters most affecting the performance of the sampling strategy.

The results obtained by the simulation model illustrated in the previous chapter provided the base on which the analysis reported in this chapter was built. Modifications to the system configuration were also considered in order to analyse the impact on the performance measures of some variables related with the system configuration, such as the number of stations in the segment and the number of machines in a station. In order to generalise the conclusions drawn for the simulated systems, a basic model, consisting of only two stations and a buffer, was also developed. Besides supporting the investigation on the validity domain of the conclusions drawn from the initial analysis, this basic model also allowed the determination of the hypothesis required for the application of the analytical models developed. For this reason, it will be presented in the next chapter.

Using the simulation approach, the analysis of the responsiveness of the sampling strategy to quality failures was also conducted. Different defect introduction modes were simulated and the sensitivity of the quality risk related performance measures with respect to some control parameters was analysed. The results obtained on the defect detectability of the sampling strategy adopted will conclude this chapter.

4.2 Experimental design

Based on the original model, different variants were developed in order to investigate the effect of some control parameters on the performances of the sampling strategy adopted, in terms of both the number of unsampled items between consecutive samples and the time between samples. Following on from discussions with the production staff familiar with the segment, three parameters were considered of primary interest in the initial analysis. They consist of the line speed, the line configuration and the sampling intervals.

The line speed is expressed here in terms of WIP-Turns, which is a measure of how fast the WIP in a line is turned over. This performance measure can be improved without recourse to reducing the processing times of stations, for example by optimising the inventory release strategy, increasing the utilisation and/or availability of existing machines and providing additional capacity at key workstations. Increasing the line speed by reducing the processing times usually proves a very costly and, in most cases, not feasible solution [5]; this is because the processing times are constrained by the technology available. As in the case of the company supporting this research, when the state-of-the-art technology is implemented in the system, reducing the processing times becomes a very difficult, if not impossible task. The increase in line speed was implemented with interventions on queuing times using two different strategies which will be illustrated later on in this section. From a quality point of view, increase in line speed will sort similar effects independently on the fact that processing times or queuing times are involved.

A variation in line speed may not directly impact the quality of the product but may impact the level of risk associated with a sampling strategy. Indeed, among production personnel of the company involved in this research, there was a strong belief that an increase in line speed would prove beneficial from a quality control point of view. This was based on the consideration that if the items cross the segment in a shorter time the quality information they carry with them will also be available more quickly.

Although the concept of line speed is intended here differently from the concept of line speed as envisioned by Inman et al. [21], there are strong commonalities between the essence of the analysis developed in this research and Inman's invitation. Indeed, the main objective of Inman's paper was to make researchers aware of unexplored fields of research in the intersection area between quality and production design issues. As the authors state, the twenty-one areas identified in the paper are definitely not exhaustive.

The line configuration refers to the station width, which is intended as the number of parallel machines operating in a station. It should be interesting to analyse what happens when the product flow is spread across a higher number of machines.

Finally, the sampling intervals of the two monitored products indicate the frequency with which the monitored items are sampled. That obviously impacts the sampled fraction of the entire production volume, provided that the production volumes of each product are kept constant.

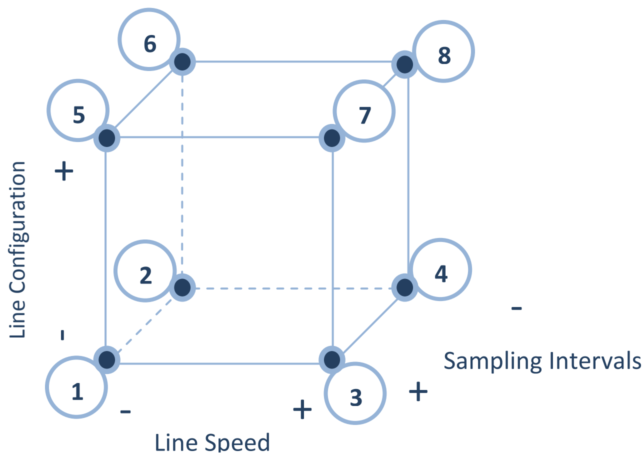


FIGURE 4.1 EXPERIMENTAL PLAN.

The experimental plan initially developed was based on a 2^3 factorial plan. Two levels for each of the three parameters were considered and all the eight combinations between them simulated. Figure 4.1 schematically represents the eight scenarios investigated; the numbers at each corner indicate the denomination with which the scenarios will be referred to hereafter.

Avoiding absolute figures for confidentiality reasons, with the line speed at a high level the time needed for the items to cross the segment is reduced by about 35% when compared with the low level.

The low level for the sampling intervals means that the intervals between two successive samples are narrower, that is the monitored items are sampled more frequently. For the low level, the sampling intervals for both the products are almost half of the sampling intervals used in the high level. Later on in this section, the reasons behind the choice of the levels will be given.

The passage from the low to the high level for the line configuration consisted of adding an extra machine in all the stations of the segment, apart from station 3. A machine was added in the stations where a reduction of the upstream routing times was considered feasible. For station 3, which shares the same working area of station 2, this reduction did not make sense.

The higher capacity obtained for the large configuration could well justify the reduction in queuing times and hence the increase in WIP-Turns. In fact, when an extra machine is available, the buffer upstream of the station where the extra machine is located is emptied in a shorter time. As a result, the queuing times were reduced; the processing times obviously do not change. This solution was supposed to be the easiest way to increase the line speed. In order to complete the factorial plan, another possible way to increase the line speed without increasing the system capacity had to be found. It involved considerations on the inventory level. In fact, since an increase in WIP-Turns can be obtained by acting on the amount of inventory kept in the system, a leaner line can reach the same objective as an enlarged configuration from a WIP-Turns point of view. In the case of a leaner line, the reduction of the routing times is obtained by reducing the number of items present in each station. This is possible by reducing the inter-arrival rate of each product. As a consequence, a scenario with the higher line speed and the initial configuration can be considered realistic. The only drawback is that, since the number of items in the segment is reduced, keeping the same sampling intervals would lead to fewer measured items. In order to allow consistent comparisons between scenarios differing for the line speed it is reasonable to maintain the same number of samples per period of time as in the initial configuration and because of this the sampling intervals for both the product types to be measured were properly adjusted. The changes to the simulation model needed to implement this scenario only required a variation of the inter-arrival and queuing and transportation time parameters which are provided to the simulation software by means of an Excel file. While for the line configuration changes a modification of the model itself was needed (Figure 4.2). In fact, from the configuration point of view the model developed proved quite rigid even though the modularity with which it was conceived made the variation of the number of parallel machines in a station quite fast.

Although not logical from a managerial point of view, for completeness sake, the lower line speed, that is the higher routing time values, in combination with the enlarged configuration was also simulated. This was done to provide the analysis reported with scientific rigours.

It is worth noting that the number of samples per unit time might differ between the scenarios simulated. It is higher when the small sampling intervals are combined with the lower line speed value; it is lower when the lower sampling intervals are combined with the higher line speed. In order to simulate configurations which would generate more samples it was assumed that the inspection station has sufficient excess capacity to manage the increased workload.

To the eight scenarios which made up the factorial plan, two more were added in order to investigate more closely the impact of the sampling station width on the monitoring ability of the sampling strategy in the other stations of the segment. This choice was motivated by the initial believe that the width of the sampling station played a fundamental role in the determination of the number of samples per unit time.

Table 4.1 summarises the characteristics of all the scenarios investigated for the first part of the analysis. The column "Global Sample Rate" refers to the global rate at which items are sampled; it is calculated as the inverse of the weighted sum of the sampling intervals adopted for each monitored product. The weights used are the production volumes of the monitored products, which can be derived from the average inter-arrival times. The global sample rate gives an indication of the sampled fraction which characterises each scenario. As will be shown in the next chapter, the sampled fraction turned out to be one of the most important parameters for determining the monitoring efficacy of the sampling strategy. The number of machines for station 4 is reported in grey, since, due to the particular nature of the process carried out, the data from this set of machines was not analysed. In fact, unlike the other stations in the segment, in station 4 two serial operations are simulated. The first operation type is performed on all the items arriving at the station; whereas, the second operation type, which in the real facility is performed in a different station, is applied only on a fraction of the items undergoing the first operation. These items are randomly chosen based on a predefined probability. The decision to group the two serial operations in one station was based on the intent to reduce the simulation model size.

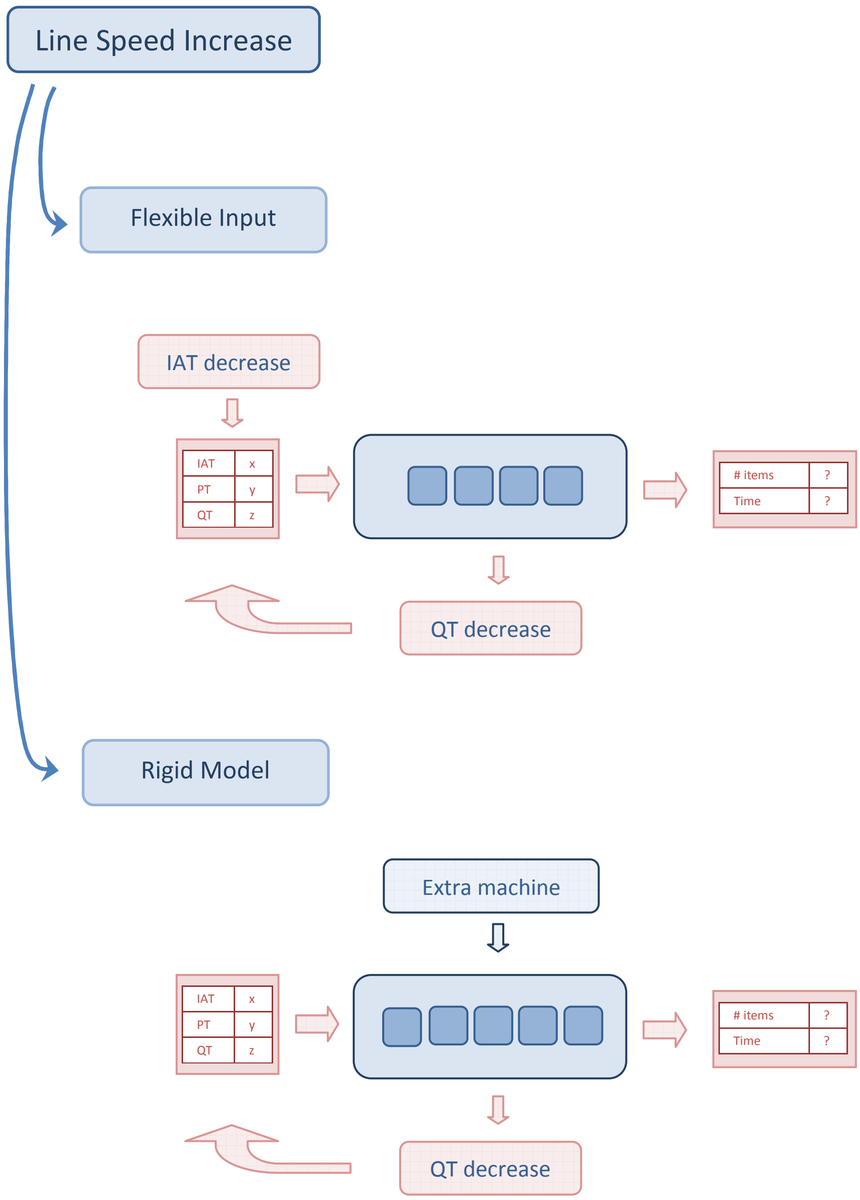


FIGURE 4.2 SCHEMATIC REPRESENTATION OF THE TWO MODALITIES USED FOR IMPLEMENTING THE LINE SPEED INCREASE.

TABLE 4.1 WIP TURN AND SAMPLING FREQUENCY FOR EACH SCENARIO SIMULATED.

Scen.	WIP turn	Number of Machines						Product Sampling Intervals		Inter-Arrival Rate		Global Sample Rate
		St 1	St 2	St 3	St 4	St 5	Insp.	A	B	A	B	
1	Low	4	4	3	3	5	4	Large	Large	High	High	Low
2	Low	4	4	3	3	5	4	Small	Small	High	High	High
3	High	4	4	3	3	5	4	Large	Large	Low	Low	Very Low
4	High	4	4	3	3	5	4	Small	Small	Low	Low	Low
5	Low	5	5	3	4	6	5	Large	Large	High	High	Low
6	Low	5	5	3	4	6	5	Small	Small	High	High	High
7	High	5	5	3	4	6	5	Large	Large	High	High	Low
8	High	5	5	3	4	6	5	Small	Small	High	High	High
9	Low	5	5	3	4	5	5	Large	Large	High	High	Low
10	Low	5	5	3	4	7	5	Large	Large	High	High	Low

4.3 Results analysis

Along with the impact on the performance measures of the parameters which vary in the experimental plan, the proximity to the inspection station was investigated. In this case, the presence of different stations progressively closer to the inspection in any of the scenarios simulated does not require any additional experiment. In fact, any single scenario could provide useful and sufficient information to conduct this particular analysis.

The analysis of the interactions between the three factors investigated is not reported since it was deemed not fundamental for an initial understanding of the effectiveness of the sampling strategy. This was based on the consideration that whilst the cause-effect relationships of the single factors on the quality risk related performance measures can be assessed without involving quantitative evaluations, the effects of the interactions are strongly dependent on the particular values of the contributing factors. In other words, a general pattern of the effects of the single factors could be extrapolated from the data available; on the contrary, the patterns of the effects of the interactions would not allow a straightforward generalisation. The way the factors interact between each other will be encapsulated in the deterministic formulae for the average values of the time between samples (Section 5.2.1) and the number of consecutive unsampled items (Section 5.2.2).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Product Type	Item ID	Mach. ID Stn 1	Mach. ID Stn 2	Mach. ID Stn 3	Mach. ID Stn 4	Mach. ID Stn 5	Mach. ID Insp.	Gen. Time	Time IN Stn 1	Time OUT Stn 1	Time IN Stn 2	Time OUT Stn 2	Time IN Stn 3	Time OUT Stn 3
2	1	7	4	3	1	3	4	1	4.537218	4.9067812	5.3938496	6.7284126	8.8500822	8.8665129	9.3020104
3	2	3	2	4	1	3	3	3	0.035977	1.3497245	2.2240466	6.7284126	7.4874663	7.5034007	8.212613
4	2	9	4	2	3	3	2	2	9.138986	9.4579875	10.096441	10.381081	11.123439	11.157173	11.70783
5	1	5	3	2	2	3	2	1	2.882501	8.4926816	9.5215762	9.5478769	10.270719	10.290678	10.734862
6	1	4	1	1	1	3	3	2	0.407429	7.8854074	9.4875426	9.73473	10.563976	10.578468	11.183762
7	2	1	1	1	2	3	5	4	0	7.8854074	10.174697	10.449045	12.270963	12.291354	12.704493
8	1	2	1	4	1	2	5	2	0	7.8854074	11.397691	11.729996	12.487556	12.498819	13.150806
9	1	8	3	4	1	3	5	0	5.954516	8.4926816	9.2514232	9.4380151	11.082798	11.100389	11.847484
10	1	6	4	1	2	3	1	3	3.046124	3.6236846	4.5535142	6.7284126	7.6585587	7.6678868	8.1268628
11	1	10	2	1	2	2	3	0	13.70669	14.287956	15.942744	16.28418	17.622656	17.64017	18.138125
12	1	11	2	2	2	2	2	0	14.74777	14.928972	16.18672	16.283455	17.290871	17.310506	17.778593
13	1	13	3	3	3	3	1	0	21.26478	21.812483	22.568208	22.788202	23.661805	23.664436	24.235059
14	1	12	1	1	3	3	2	0	16.50637	16.713726	17.362189	17.450975	18.534689	18.546161	19.00878
15	1	14	3	4	1	3	4	0	23.7779	24.11051	25.062907	25.225456	25.992935	26.008777	26.548775
16	1	17	4	1	2	1	4	0	28.59507	28.948563	29.245823	29.63253	30.584464	30.5976	31.21162
17	1	15	3	1	2	1	2	3	24.00244	24.477253	25.148138	25.435373	26.840682	26.854451	27.507772
18	2	18	2	3	1	3	4	2	30.23635	30.444519	31.863226	32.07049	33.306719	33.320869	33.7666

FIGURE 4.3 TYPICAL EXCEL OUTCOME AT THE END OF A SIMULATION RUN.

In order to avoid unnecessary complications to the simulation model and reduce the simulation run times, the two performance measures on which the analysis was based are not a direct output of the simulations. In fact, as mentioned before, at the end of each simulation run, an Excel file containing all of the information gathered with respect to the items processed in the segment is generated. This file is composed of four worksheets, one for the serial flow items and the other ones for the cross flow items. These sheets are similarly structured (Figure 4.3). The first column indicates the product type; three id's are used reflecting the three categories identified in the system description. The second column contains an item id, which is applied exclusively to the serial flow items and represents the order with which those items are generated within the segment. From the third to the eighth column the id's of the machines visited by the items in each of the six stations in the segment are registered. The timestamps relative to the item generation and the times in and out of each visited machine are reported in the remaining thirteen columns. Each item corresponds with a row in a worksheet and whenever the information relative to a column is not applicable, the corresponding cell is left blank. Keeping the same structure for all the worksheets, independently of the particular product type, eased the output data processing which was conducted using the software Matlab v7.1®. Different Matlab functions were developed in order to calculate the number of unsampled items between consecutive samples and the time between samples. Different versions of those functions are available; they try to respond to the different objectives of the analyses conducted. One of the functions used is shown in Appendix C.

The simulation model was run for 6,000 hours with a warm-up period of 1500 hours and 5 replications were conducted with the model each time an experimental scenario was investigated. Data from the models was averaged across the 5 simulation runs. This provided a population of about 3,000 samples behind each reported statistic.

Sections 4.3.1 to 4.3.4 will focus on the impact of each of the factors investigated on the quality risk related performance measures, namely the line speed, the sampling interval, the proximity to the inspection station and the station width. A table summarising the main findings of these analyses is reported in Section 4.3.5.

4.3.1 Line speed impact

At the beginning of the analysis great interest was paid by the production staff involved in this research to the impact that the line speed could have on the sampling strategy performances. The reason is that the predictions of the effects it could have on the monitoring speed were not straightforward and the opinions each person had were different. However, the feeling that speeding up the line could have beneficial effects on the monitoring speed was definitely more common. If the items went faster through the line, in the same timeframe, more samples could be measured and the information about the quality status of the line could be updated more frequently. That could result in a fewer number of items exposed to the risk of being produced by an out of control machine.

The results obtained from the simulations seem to contradict this opinion. Figure 4.4 shows the patterns for the mean and some high percentiles of the distribution of the number of unsampled items between consecutive samples based on the results obtained for the second station. The sequence of scenarios on the horizontal axis is reordered so that a direct comparison between scenarios differing exclusively for the line speed is easier. The similar patterns of the mean and the percentiles up to the 95th percentile should guarantee that any consideration made about the mean values is still valid for the whole distributions.

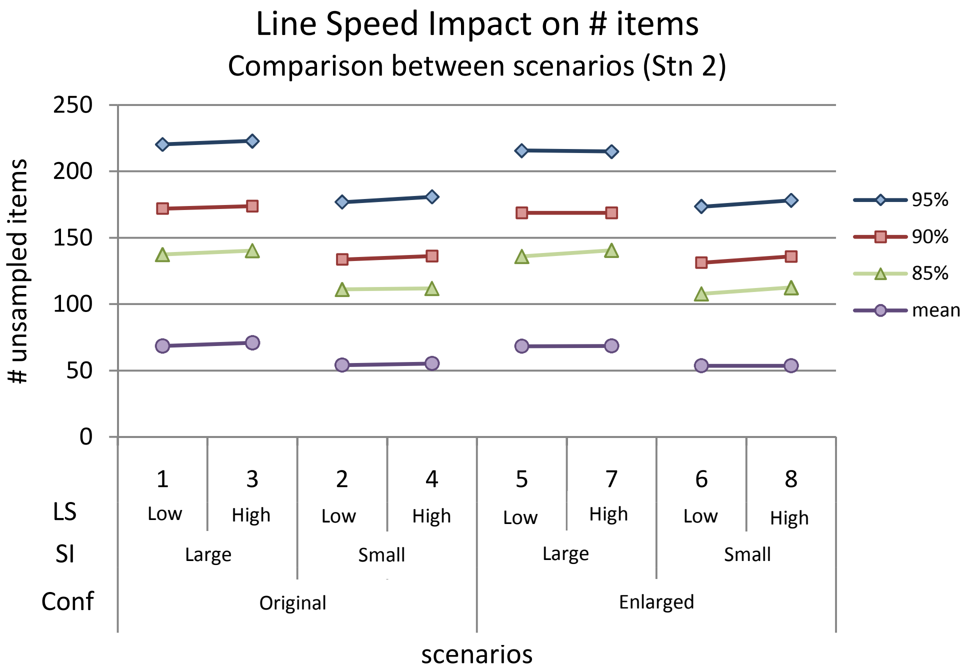


FIGURE 4.4 IMPACT OF THE LINE SPEED ON THE NUMBER OF UNSAMPLED ITEMS BETWEEN CONSECUTIVE SAMPLES.

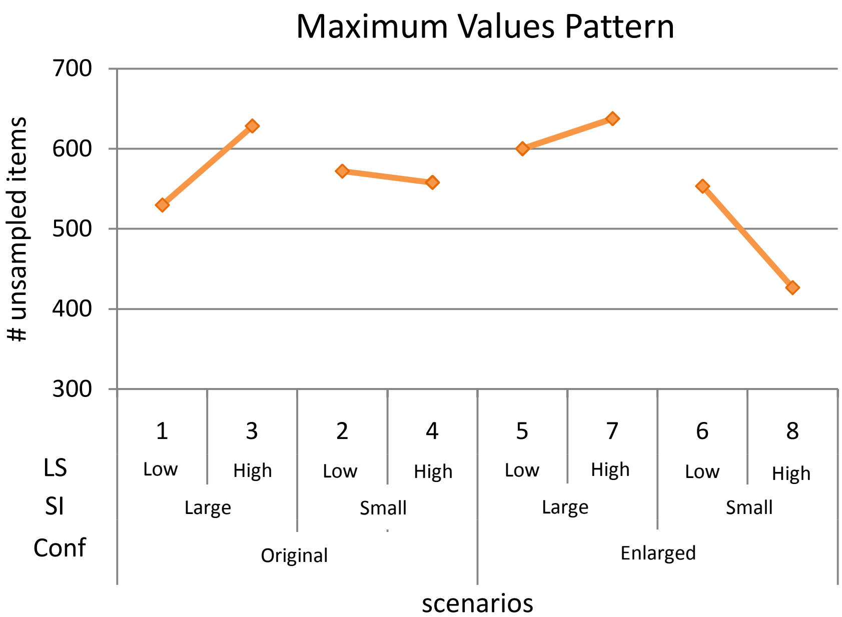


FIGURE 4.5 PATTERN OF THE MAXIMUM VALUES OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN CONSECUTIVE SAMPLES.

Any eventual variation of the pattern of the maximum values should be more likely due to anomalies rather than to a systematic behaviour of the system. For demonstrative purposes, the pattern of the maximum values obtained for the same station as the one considered in Figure 4.4 is shown in Figure 4.5. Strong anomalies can be noted and for this reason the maximum values will never be taken into account in this analysis hereafter. The exact same observations apply to the other performance measure considered, the time between samples, which is shown in Figure 4.6.

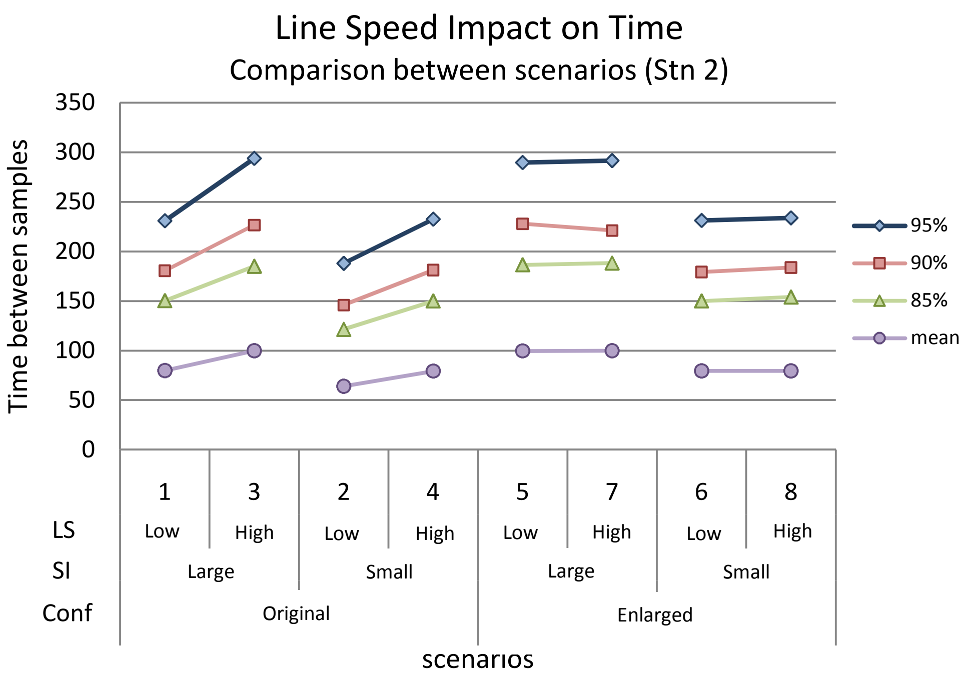


FIGURE 4.6 IMPACT OF THE LINE SPEED ON THE TIME BETWEEN SAMPLES.

Figure 4.4 indicates that the number of consecutive unsampled items for each machine, which was supposed to decrease by increasing the WIP-Turns, appears not to be affected by the line speed when the sampling rate is kept constant in both the configurations. In fact, the variation of the line speed affects all the product flows in the same measures; this means that the volumes of the different product categories are kept unchanged. As a consequence, the increased line speed can not be noticed on the number of items; however, this will most likely impact the time between samples as, if the line speed is higher, this should mean that the same quantity of items is produced in a shorter time, at least for the case when the inter-arrival times are affected by line speed variations.

This supposition is contradicted by the results reported in Figure 4.6. In fact, for the enlarged configurations, the time between samples seems to be independent of the WIP-Turns and it even increases with the line speed, for the original configurations. The explanation of this behaviour is found if the line speed variation is analysed in terms of the impact it has on the inter-arrival time at a machine level. In the original configuration, the line speed increase is actually obtained by increasing the inter-arrival time at a station level. Since the number of machines for these scenarios (scenarios 3 and 4) does not change with respect to the corresponding low speed scenarios (scenarios 1 and 2), the increase in the inter-arrival time implemented at a station level means that the inter-arrival time at a machine level proportionally increases. The increased time between samples is a consequence of the slower time at which items moves in and out of the segment. In the enlarged configurations, an increase in the inter-arrival time is still observed at a machine level. In this case, it is caused by the presence of an extra machine in the stations and not by a variation of the inter-arrival time at a station level. The increased inter-arrival time at a machine level causes the increase in the time between samples for all the scenarios with the enlarged configuration in comparison with the scenarios using the original configuration. In particular, just by coincidence, the variation that the extra machine in the second station determines on the inter-arrival times for scenario 5 (and 6) is the same as the one imposed in scenario 3 (and 4); hence, the same results. However, looking at the enlarged configuration scenarios, the passage from the low line speed to the high line speed (scenarios 5 to 7, and 6 to 8) does not have any impact on the time between samples. In fact, the speed was increased by making the production line leaner, that is, reducing the queuing times. The inter-arrival times between the low line speed scenarios and the high line speed scenarios remain unchanged; that means no variation is registered for the times since no variation was detected in the number of unsampled items between samples.

The same effect was seen across all the stations in the segment. The results shown in Figure 4.4, Figure 4.5 and Figure 4.6 report results normalised with respect to results obtained in the fifth station for the first scenario. This was done for confidentiality reasons and, where not differently specified, it will hold for all the figures hereafter.

In conclusion, matching the number of samples per unit time, as happens for scenarios 1 and 5, the risk associated with the system quality performances theoretically doesn't change, in terms of number of unsampled items between samples (Figure 4.4). However, in the enlarged configuration, which appeared to be the simplest implementation of the decision to speed up the line, the time between samples significantly increases (Figure 4.6). It is worth noting that the presence of an extra machine in each station is by itself a cause of further risk and costs.

4.3.2 Impact of changing the sampling interval

When everything else is kept constant, the variation of the sampling intervals effectively means the variation of the sampled fraction. In particular, if the frequency with which items are sampled increases it should be reasonable to expect a reduction of the quality risk, which means a reduction of both the performance measures.

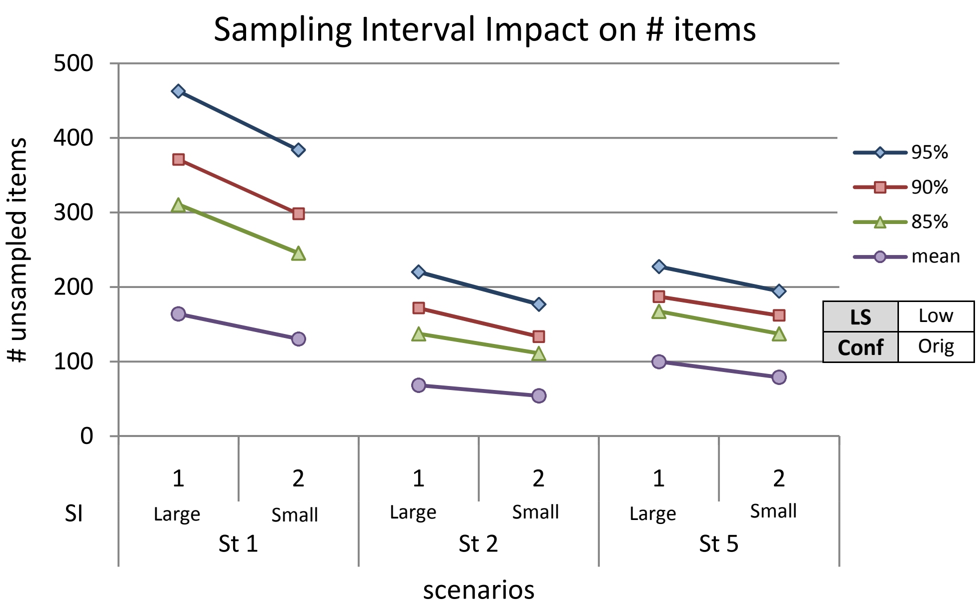


FIGURE 4.7 IMPACT OF THE SAMPLING INTERVAL ON THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS IN THREE STATIONS OF THE SEGMENT.

The results shown in Figure 4.7 confirm this reasoning. The first two scenarios are compared in terms of number of unsampled items between consecutive samples per machine for three stations in the segment.

Independently of the station, the number of unsampled items significantly decreases when the sampling intervals for both the monitored products decrease. The slope of the segments represented in Figure 4.7 is very similar for the different stations. It approximately equals the variation of the global sampling intervals.

In order to observe a beneficial effect on the number of consecutive unsampled items, the global sampled fraction and not only the sampled fraction of the monitored products has to increase. Indeed, if the sampled fraction increase is implemented with respect to the monitored volume, that is smaller sampling intervals are set for the monitored products, it can happen that a contemporary increase in the unmonitored volume crossing a station could determine a reduction of the global sampled fraction with consequent deleterious effects on the number of unsampled items.

The different quantities of consecutive unsampled items which characterise the different stations are due to the different unmonitored volumes processed. The only other parameter which could potentially be the reason for the reported differences is the station width. It is worth noting that both the variation of the unmonitored volume and the different width determines different loadings of the machines in terms of the global number of processed items. This means, the variation of the volume processed by each machine, independently of the phenomena by which it is caused, is the ultimate reason for the differences between the number of consecutive unsampled items observed in the different stations (Figure 4.7).

Since there is no impact from the unmonitored flow on the number of unsampled items between consecutive samples, any difference in the time between samples should be related to the number of machines in the station. The effect of station width may be examined by considering the results from stations 1 and 2 (same number of machines) and station 5 which has a larger number of machines. The first and the second stations show the same results while the fifth station is characterised by higher times (Figure 4.8) This is reasonable since the number of monitored type items is the same through the stations; the wider the station is, the more spread is the product flow between the machines and the greater is the time between two successive samples.

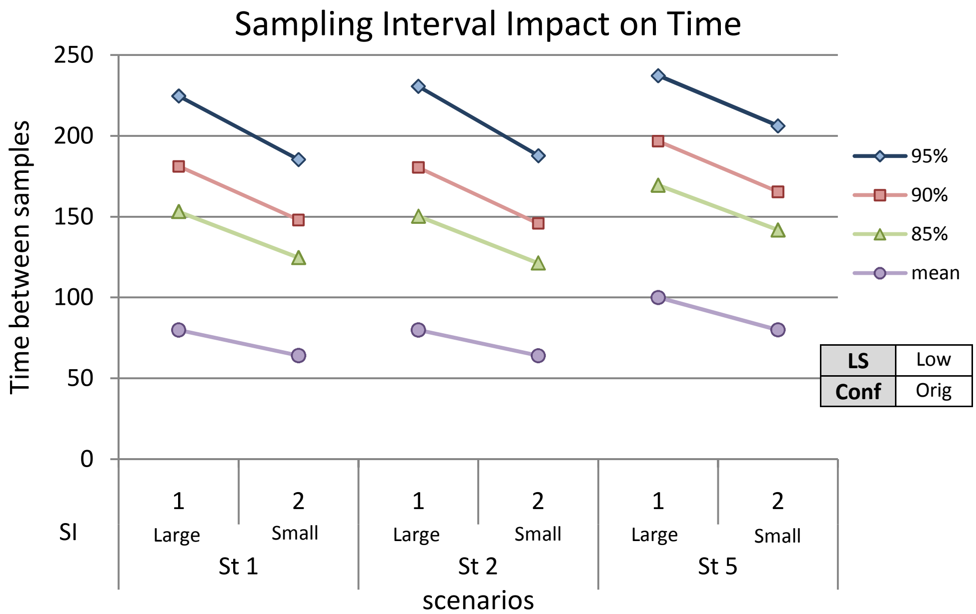


FIGURE 4.8 IMPACT OF THE SAMPLING INTERVAL ON THE TIME BETWEEN CONSECUTIVE SAMPLES IN THREE STATION OF THE SEGMENT.

As happens for the number of unsampled items, the magnitude of the reduction of the time between samples caused by the increased sampled fraction is similar for the three stations represented and reflects the average variation of the global sampling interval.

Considering that, in relative terms, the impact of the sampling interval on both the performance measures is similar in the different stations, further analyses can focus on one station only. The first station has been chosen as a reference. The comparison between the number of consecutive unsampled items between samples is proposed in Figure 4.9. Figure 4.10 shows the impact of the sampling interval on the time between samples. In both the figures, the data are grouped so that a direct comparison is allowed for scenarios differing only for the sampling intervals.

Independent of both the configuration and the line speed, the main observation is still that a decrease of the global sampling interval causes a relevant reduction of both the performance measures.

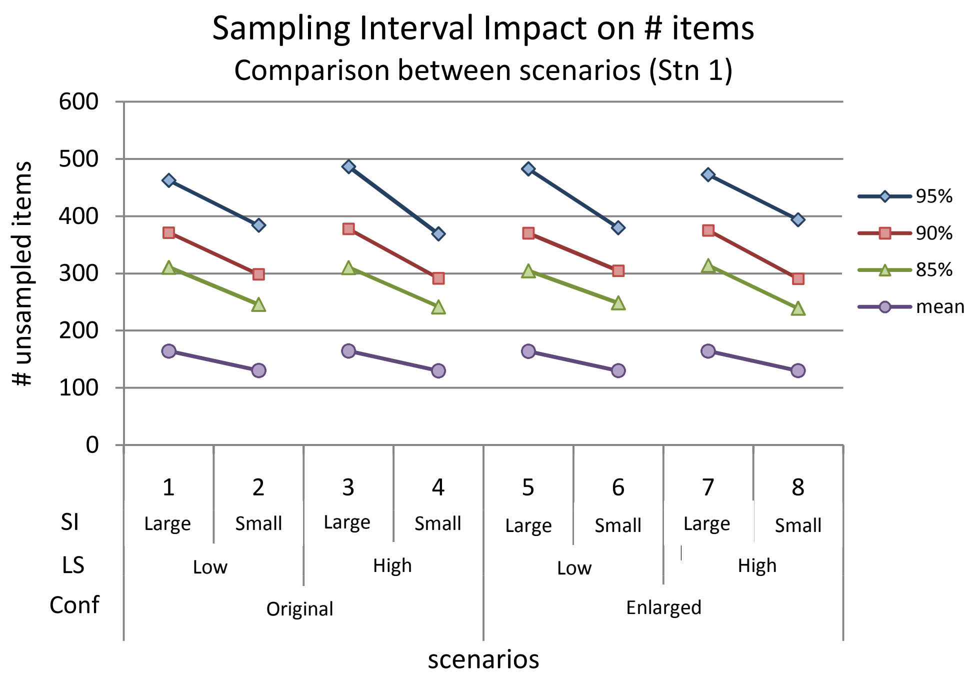


FIGURE 4.9 IMPACT OF THE SAMPLING INTERVALS ON THE NUMBER OF ITEMS BETWEEN CONSECUTIVE SAMPLES IN DIFFERENT SCENARIOS.

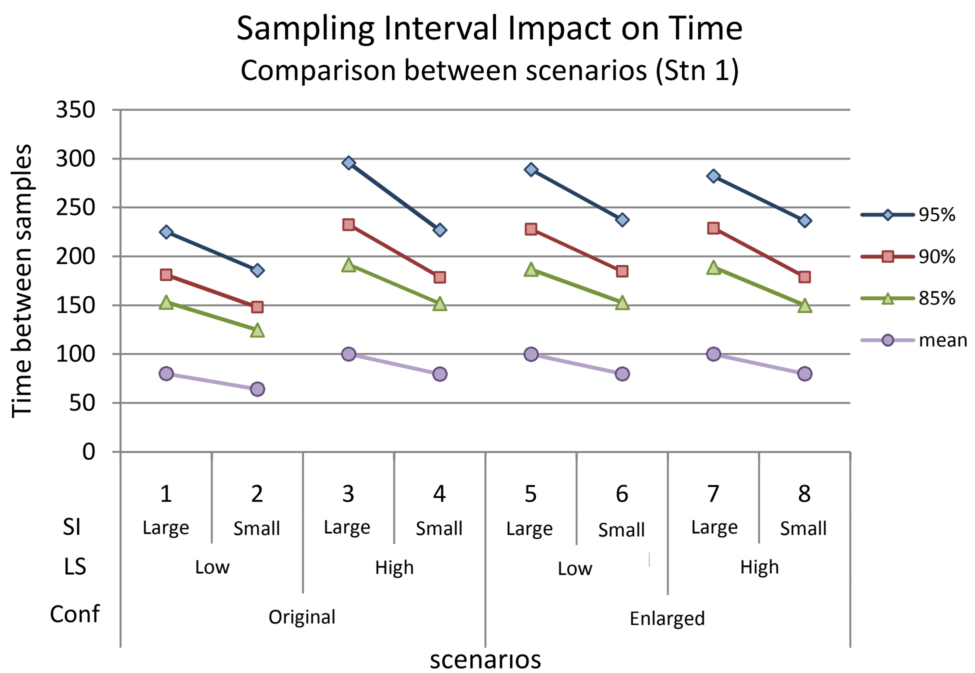


FIGURE 4.10 IMPACT OF THE SAMPLING INTERVALS ON THE TIME BETWEEN SAMPLES IN DIFFERENT SCENARIOS.

The magnitude of time reduction between the grouped scenarios is independent of the line configuration, i.e. the station width, and reflects the average increase in the number of samples. The higher values of the time between samples for scenarios 3 and 4 with respect to scenarios 1 and 2 are a consequence of the reduced inter-arrival rate for all the products, and in particular for the monitored products. In fact, if the sampled fraction is kept constant but the frequency with which all the products arrive is lower, it is understandable that nothing changes in terms of the number of unsampled items between samples (scenarios 1 and 3 in Figure 4.9) but the time needed to observe two consecutive samples increases.

In conclusion, a reduction of the sampling intervals of the monitored products reduces the quality risk in that it results in an increase in the sampled fraction. In the case where the sampling intervals reduction is followed by an increase of the unmonitored flow the effects may be balanced, which means the beneficial effects of the sampling intervals can be apparently reduced or even lost since the sampled fraction of the global volume might, as a result, be decreased. At the end of this analysis, it is clear that talking in terms of global sampled fraction allows a more immediate evaluation of the impact that the sampling intervals of the monitored products have on the quality risk performance measures, and in particular on the number of unsampled items between consecutive samples. The reason is that the global sampled fraction is a more comprehensive measure, which includes in itself information about the production volumes of the different product categories.

4.3.3 Proximity to the inspection station

The proximity to the inspection station is another factor that logically could affect the monitoring capability of the system, and in particular, the speed of defect detection and feedback information. This is investigated here in terms of the impact on the number of unsampled items between samples and the time between samples. These measures quantify in relative terms the delay in quality failure detection for a particular machine. In fact, they take as a reference the processing of the previous sample in that machine and ignore the number of items processed during the time needed for the sample to reach the inspection station. On the other hand, investigating the impact of the proximity to the inspection station from an absolute perspective can be trivial; it is understandable that the further a station is from the inspection station the later the feedback information will be available. The same can't be said for the relative perspective.

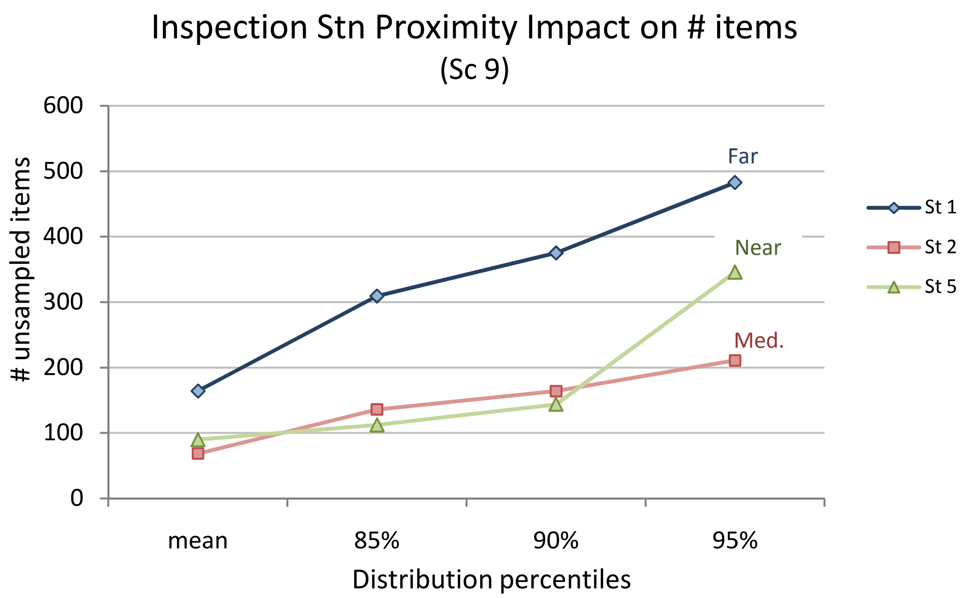


FIGURE 4.11 MEAN AND SOME PERCENTILES OF THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS FOR STATIONS 1, 2 AND 5 (SCENARIO 9).

Amongst the scenarios simulated, scenario 9 offers the best opportunity to investigate the impact of the proximity to the inspection station since it is the only one which contains three different stations of the same width. Focusing on these three stations avoids the distortions that the differences between station widths can cause to the analysis. In Figure 4.11, the comparison between the stations proves very difficult; the different values of the mean are more likely caused by the different cross flow volumes. It is worth noting once again that since the cross flow is independent from station to station, the unmonitored volumes processed in each station varies and, as a consequence, the number of items between samples, which depends on the global production volume, does not necessarily provide useful indications for this analysis purpose. The proximity of the station to the inspection station can play a marginal role too in the difference between the mean values in Figure 4.11. For clarity, station 1 is the farthest from and station 5 the closest to the inspection station. The high value of the 95th percentile of the distribution of the number of unsampled items for the fifth station is probably due to an anomalous dispersion of the simulation results.

Considerations based on the time between samples should be more reliable. In fact, Figure 4.12 shows more regular patterns for some statistics of the distribution of the time between samples for the three stations. The mean values do not change; that implies that the differences in the means of the number of items between samples are only due to the different unmonitored volumes processed in each station. As regards the higher percentiles, Figure 4.12 indicates an interesting reduction of the slope of the lines as the inspection station gets closer. Whereas the relevantly different dispersion of the distributions between stations 1 and 2, and station 5 is due to the fact that station 5 is a sampling station, the slight reduction of the line slope observed between station 1 and station 2 reveals that also the proximity to the inspection has a positive effect on the variability of the distributions.

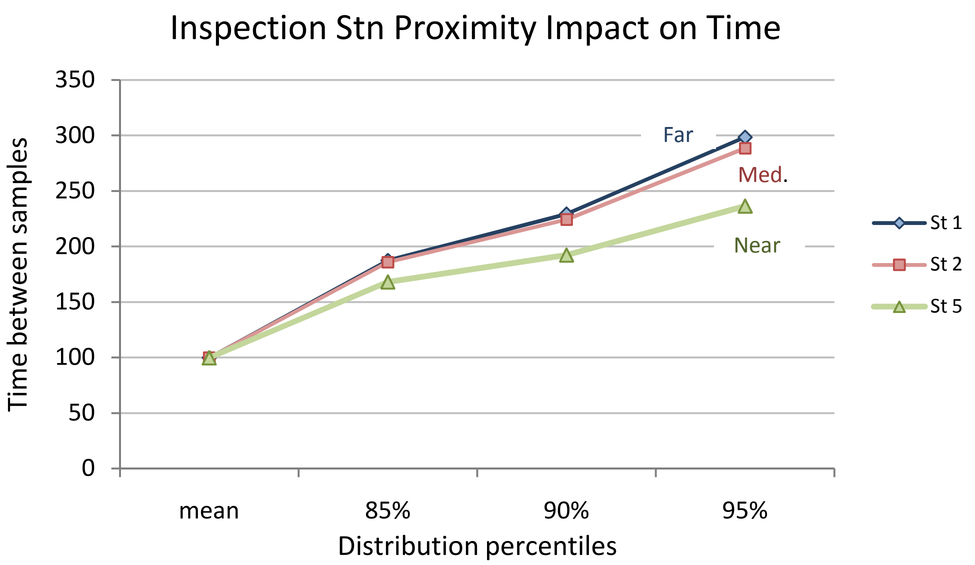


FIGURE 4.12 MEAN AND SOME PERCENTILES OF THE DISTRIBUTION OF THE TIME BETWEEN SAMPLES FOR STATIONS 1, 2 AND 5 IN SCENARIO 9.

In station 5, which is the sampling station, the time graph presents very low upper percentiles. That can be seen as the effect of the deterministic sampling intervals. The higher regularity which characterises this station is even more evident in Figure 4.11, where, up to the 90th percentile, the trend of the line representing station 5 is flatter when compared to the other stations. The lack of any deterministic component and the distance from the sampling station determine a higher variability in the other stations for both the performance measures. It has to be noted that the different variability of the inter-arrival times for the cross flow also contributes to the variability of the number of items between samples.

The involvement of a deterministic factor in the sampling station has diverted the focus of this analysis from the inspection station itself to the sampling station. In fact, the results shown are more indicative of the impact of the distance from the sampling station rather than from the inspection station. However, in most cases, the sampling and the inspection stations are immediately close, as happens in this system.

In the end, the distance from the sampling station, contrary to what one may expect, does not have a remarkably negative effect on the monitoring capacity apart from a slight reduction of the variability which is detectable for both the measures.

4.3.4 Station width

Due to its impact on the inter-arrival time at a machine level, the station width is expected to be a relevant factor for the effectiveness of the sampling strategy, in particular from a time perspective. In fact, keeping the inter-arrival time at a station level constant, the wider the station is, the higher the inter-arrival time of the items at each machine will be. This immediately implies that more time is needed for a machine to receive the same quantity of items and, as a consequence, the arrival of a sample will be delayed. The fact that the time is expanded should not necessarily have an impact on the number of items between samples.

The comparison between the results of scenarios 2 and 6 provides a better understanding of the nature of the impact of the station width on the performance measures. In fact, the scenarios only differ for the line configuration, i.e. the number of machines in the stations. Apart from station 3, which operates with only three machines for both the scenarios, all the other stations in scenario 6 contain an extra machine with regards to the original configuration in scenario 2. The production volumes in each station and the sampling intervals of the monitored product types do not change between the two scenarios; hence, the resulting global sampled fraction is unchanged.

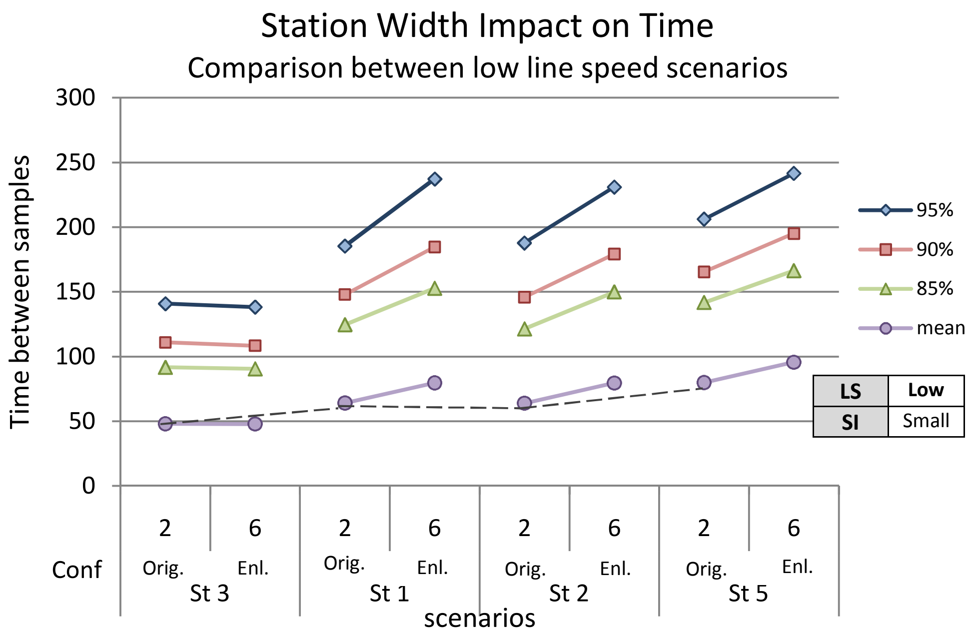


FIGURE 4.13 IMPACT OF THE STATION WIDTH ON THE TIME BETWEEN SAMPLES.

As the dashed line in Figure 4.13 reveals the time between samples increases as the number of machines increases within the same scenario; that is a direct consequence of the reduction of machine loading due to the spreading of the flow across a greater number of machines. Stations 1 and 2, which have the same width, present no difference in the time between samples. Evidently, the reason is that the inter-arrival time at a machine level for the monitored flow is the same for both the stations. The comparison of the results of the two scenarios in Figure 4.13, confirms the negative effect of the number of machines in a station on the time between samples. Adding an extra machine causes an increase in the time between samples. Independently of the station, when the number of machines matches between the two scenarios, as happens for station 5 in scenario 2 and stations 1 and 2 in scenario 6, the same times between samples are obtained. The same happens for station 3, which has the same width in both the scenarios. When an extra machine is added in a station, the differences in the time between samples between the same stations in the original and enlarged scenarios prove proportional to the variation of the inter-arrival times in each machine.

It is, therefore, clear that the time between samples is governed by the inter-arrival time of the monitored products and their sampling intervals; the unmonitored flow has no effect on it.

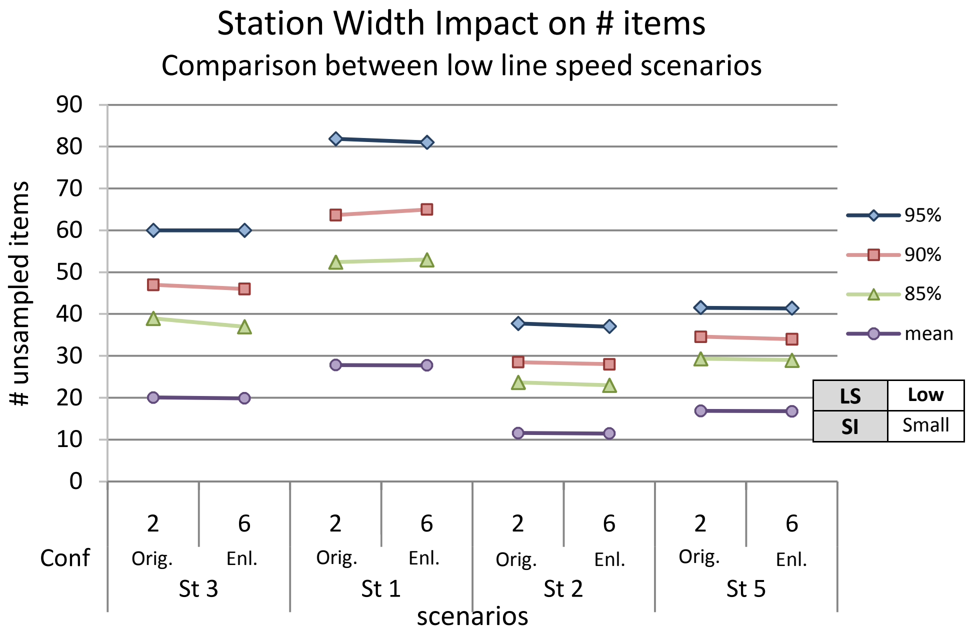


FIGURE 4.14 IMPACT OF THE STATION WIDTH ON THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS.

Due to the presence of the cross flow, the only valid comparison for the number of items between samples is limited to each single station in the two scenarios. The variation of the number of machines has no impact on this measure (Figure 4.14) and the reason for that is that the extra machine does not cause any change in the sampled fraction. The expansion of the inter-arrival times in the stations with the extra machine affects the different flows in the same proportion. For example, when the number of machines goes from 4 to 5, the inter-arrival time at a machine level increases by 5% for each of the three product flows. That means, in terms of relative volumes, nothing changes. If the sampling intervals and the proportion between the monitored flow and the unmonitored flow are kept unchanged, the number of items between samples does not change. It is interesting to notice that not even the high percentiles of the distributions change; hence, even the variability of the number of unsampled items is unaffected by the station width.

The importance of the station width on the time between samples has its origin in the impact it has on the inter-arrival time at a machine level. In the case that variations of different parameters keep the inter-arrival time unchanged, the station width apparently loses its impact. The comparison between scenarios 3 and 7 can prove this statement. These two scenarios differ for the line configuration and the inter-arrival rate at a station level. In scenario 3 the inter-arrival time at a station level for all the product flows is 20% higher than in scenario 7. Given that the sampling intervals of the two products are equal for the two scenarios, the variation in the number of samples per unit time reflects the variation of the inter-arrival rates at a station level; that means scenario 3 has 20% fewer samples per unit time than scenario 7. However, in relative terms, since the inter-arrival rate variation interests all the flows, the sampled fraction remains unchanged.

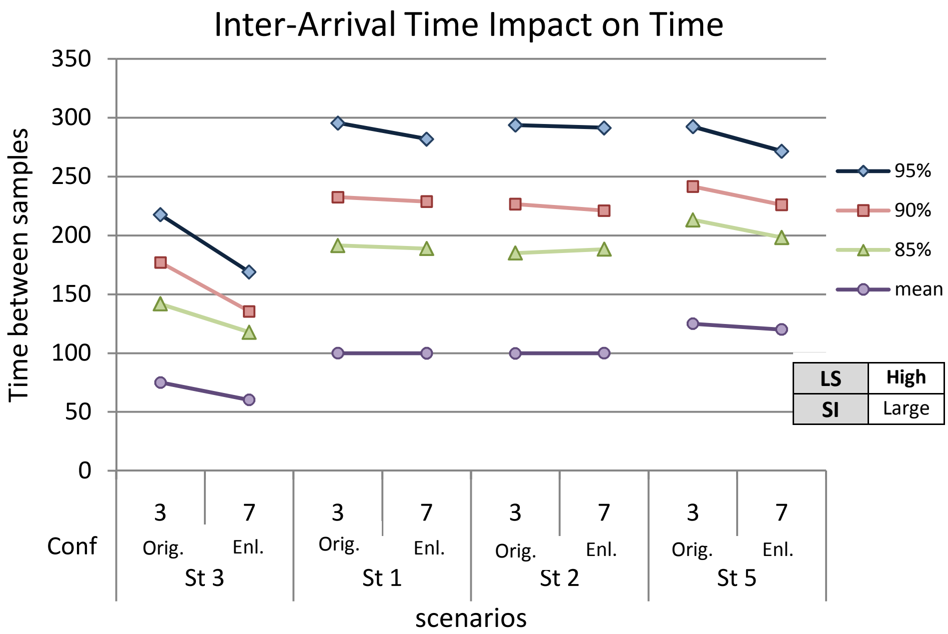


FIGURE 4.15 COMPARISON OF THE TIME BETWEEN SAMPLES BETWEEN SCENARIOS WITH DIFFERENT INTER-ARRIVAL TIMES AND DIFFERENT LINE CONFIGURATIONS.

For stations 1 and 2, the presence of an extra machine in scenario 7 causes a 20% reduction of the inter-arrival rate at a machine level with respect to scenario 3. This reduction perfectly compensates for the inter-arrival time increase imposed at a station level in scenario 7. As a consequence, the inter-arrival rate at the machines of stations 1 and 2 is the same in both scenarios. In fact, in Figure 4.15, no difference is reported for the time between samples. The reduction of the time between samples for the machines of station 3 in scenario 7 is due to the higher inter-arrival rate in scenario 7, given the number of machines remained constant. Finally, the slight difference in the time between samples for the fifth station is an effect of the change from 5 to 6 machines between scenario 3 and 7 which means the inter-arrival rate is decreased by approximately 17% in scenario 7, compared with the 20% reduction in scenario 3. This proves that as long as the sampling intervals and the inter-arrival rate at each machine are kept constant, the station width has no impact on the time between samples.

Finally, the impact of the sampling station width on the quality risk measures, when the other stations keep the same width, was investigated. Considering scenario 5 as a reference scenario, the width of station 5 was increased and decreased by a machine in scenarios 9 and 10 respectively. Since nothing changes apart from the sampling station width, the sampled fraction does not vary between the three scenarios. As a consequence, as can be seen in Figure 4.16, there was no variation of the number of items between consecutive samples in the different stations of the segment across the three scenarios. The inter-arrival time at a machine level is modified only in the sampling station; it linearly increases with the number of machines in the station. As a result, the time between samples in the sampling station increases with the same linear pattern (Figure 4.17). As was expected, the time between samples in the other stations was not affected by the sampling station width.

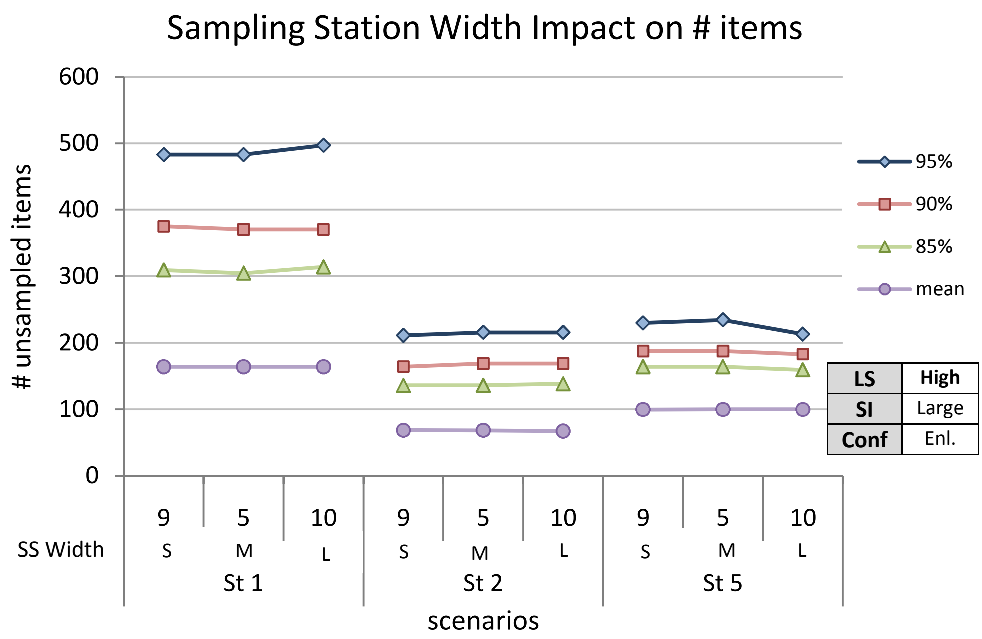


FIGURE 4.16 IMPACT OF THE SAMPLING STATION WIDTH ON THE NUMBER OF ITEMS BETWEEN CONSECUTIVE SAMPLES IN DIFFERENT STATIONS.

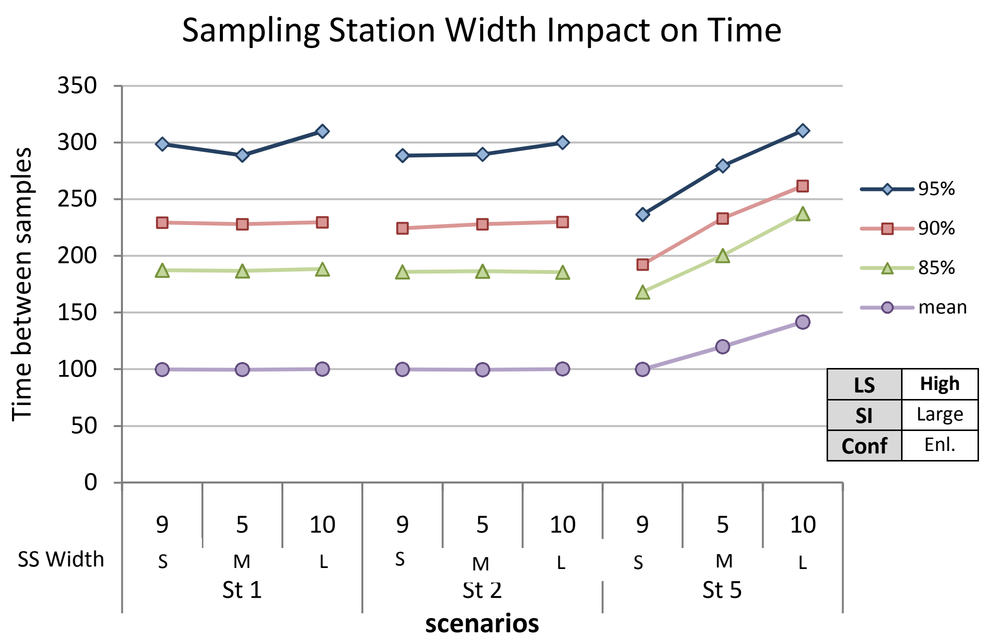


FIGURE 4.17 IMPACT OF THE SAMPLING STATION WIDTH ON THE TIME BETWEEN SAMPLES IN DIFFERENT STATIONS.

4.3.5 Results overview

The main findings of the first part of the analysis are summarised in Table 4.2.

TABLE 4.2 MAIN FINDINGS OF SECTION 4.3.

Parameter	# consecutive unsampled items	Time between samples	Comments
Line Speed	No impact	No impact for the enlarged configurations; Increases with the line speed for the original configurations	In the original configurations, a line speed increase is obtained through an inter-arrival time decrease at a station/machine level; in the enlarged configurations the inter-arrival time at a machine level is not affected by the line speed increase.
Sampling Interval	Reduces with lower sampling intervals	Reduces with lower sampling intervals	Sampling interval variations affect the performance measures when the global sampled fraction is modified by them.
Proximity to the inspection station	Irrelevant impact	Irrelevant impact	The proximity to the sampling/inspection station tends to reduce the dispersion of the performance measures. This is particularly evident for the time between samples; the number of consecutive unsampled items variability is also affected by the variability of the inter-arrival time of the cross flow products.
Station Width	No impact	Increases with the station width	The variation of the number of machines in a station causes a variation of the inter-arrival time at a machine level, hence, the impact on the time between samples. The global sampled fraction is not affected by this parameter and so the number of consecutive unsampled items.

4.4 Defect introduction

In order to investigate the responsiveness of the sampling policy in detecting the production of poor quality items, the introduction of defects in some stations of the segment modelled was simulated. The model used for this analysis purpose is the same used in the previous study; the actual configuration of the line corresponds with the configuration used in scenario 1 and the production parameters were unchanged. The stations chosen for introducing defects are the farthest one from the monitoring station, which is obviously station 1, the narrowest station in the segment, station 3, and the sampling station, station 5. Due to the sampling policy, which aims to monitor each single machine independently in all the stations, it was decided that the production of poor quality items could happen at all the machines in a station at the same time. Besides being a possible natural behaviour of the system, this choice reduces the experimental time since a sufficiently high number of samples are available in a small number of simulation runs.

Different modalities of defect introduction were considered; a malfunctioning machine may produce poor quality items in a persistent or in an intermittent fashion. The persistent mode implies that once the machine has entered a poor quality state it will only produce defective items. On the other hand, defect introduction in the intermittent mode implies that the machine produces poor quality items on a probabilistic basis and that while in this state the machine will still produce a certain percentage of items which are defect free. Following the logic of when poor quality items are produced while the machine is working perfectly, the percentage of good items produced reflects both the statistical dispersion of the production of good items and the choice of the control limits for monitoring the process. Considering what happens in the real segment, which has been modelled, another method of defect introduction was simulated. This consists of introducing defects in a permanent fashion until a repair event with duration higher than a threshold value is conducted on the machine; after this event the production of poor quality items will prove intermittent. This event, which can be usually classified as a PM event, is not intended to fix the machine. It only consists of an ordinary intervention on the machine which is triggered not by the quality failure but for reaching a due timestamp or number of completed operations. Hence, it might happen that the quality failure is not noticed during the maintenance operation and the machine is not properly fixed so that the quality status of the items produced after the reparation can be still poor based on a given probability.

Moreover, a case was studied to check how robust the sampling policy was in dealing with Type II errors. In this case, even though the machines operate in a quality control, due to the natural variability of the process, the production of good items is spaced out with the production of bad ones.

Finally, the effects of the variation of the WECO rules adopted in the sampling policy on the number of poor quality items produced were investigated. In this study, the current rule, which suggests a repair intervention when four out of five lots coming from a machine are classified as defective, was compared with a stricter rule which triggers the machine shut down when two out of three items fail the inspection.

In all the scenarios investigated the inspection is considered reliable. However, the consequences of an item misjudgement were analysed for certain values of the Type I error.

The performance measures on which the comparison of the different scenarios was based are the quantity of poor quality items produced up to the quality failure detection, which is expressed in both absolute and relative terms with respect to the whole production, the corresponding production time and the number of good items produced after the machine goes out of control. This last measure has obviously less importance than the other two for the aim of this study. The number of poor quality items produced up to the quality failure detection and the corresponding production time were conceived as the adaptation of the measures used during the first part of the analysis to the scenarios simulating the defect introduction. All the measures were calculated by processing the simulation output by means of the functions developed and presented in Appendix C.

Table 4.3 summarises the scenarios simulated for the defect detectability analysis. The initials displayed in the table will be used as a reference in the rest of this chapter.

TABLE 4.3 DENOMINATION OF THE SCENARIOS SIMULATING THE DEFECTS INTRODUCTION.

		Defect Introduction Mode				
		Persistent	**Intermittent**	**Persistent with PM events**	**WECO rule (2/3) Intermittent**	**Type II defects**
Out of Control Stn	**St 1**	D1	D2	-	-	-
	St 3	D3	D4	-	-	-
	St 5	D5	D6	D7	D8	D9

Each scenario was replicated 100 times. A terminating condition was imposed so that each simulation run was ended when the WECO rule adopted for the corresponding scenario was triggered. The simulation output was averaged across all the replications. This provided 100 samples behind each reported statistics.

4.4.1 Intermittent vs. persistent defect introduction

In order to compare the impact of the persistent and the intermittent introduction of defects on the ability of detecting the quality inefficacy, defects are introduced within the same station so that no other factor can influence the analysis. In the intermittent mode, the good quality items represent 5% of the whole machine production after the quality failure. The number of poor quality items produced before the machine shut down is shown in Figure 4.18. The event which determines the interruption of the machine operating status is the verification of the adopted WECO rule, which is four out of five consecutive samples fail the inspection. As for all the results shown in the rest of this chapter, the results in Figure 4.18 were normalised with regard to the maximum number of poor quality items produced in station 1, in scenario D1. The 95th percentile of the distribution obtained by the simulation results, the mean and the 5th percentile are shown. The choice to consider the two percentiles instead of the maximum and the minimum values is based on the consideration that the extreme values usually express anomalies, whereas the high and low percentiles are more expressive of the shape of the distributions. The production time of poor quality items, normalised with respect to station 1 results in scenario D1, is shown in Figure 4.19.

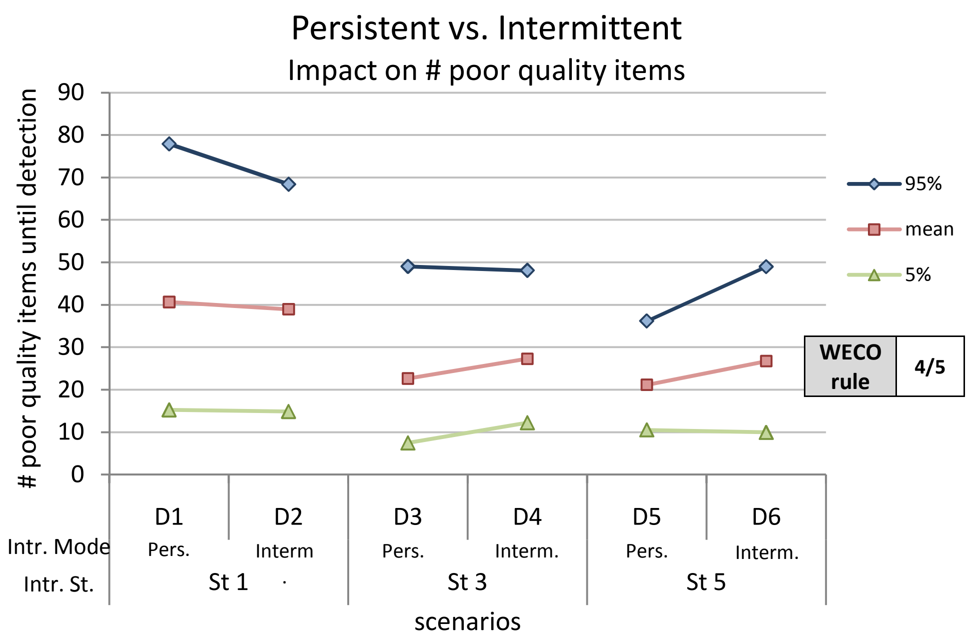


FIGURE 4.18 NUMBER OF POOR QUALITY ITEMS: COMPARISON BETWEEN THE PERSISTENT AND INTERMITTENT DEFECT INTRODUCTION IN THREE STATIONS.

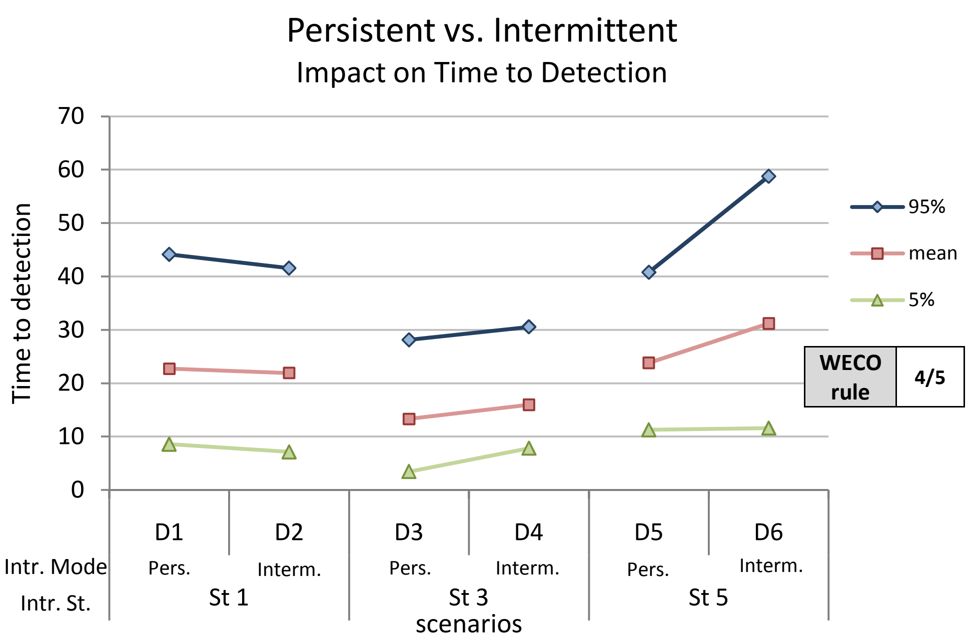


FIGURE 4.19 TIME TO DETECTION: COMPARISON BETWEEN THE PERSISTENT AND INTERMITTENT DEFECT INTRODUCTION IN THREE STATIONS.

The number of poor quality items and the time until detection present the same patterns, being the number of items produced strictly related with the production time. Apart from the machines in the first station, when the defect introduction is intermittent a delay effect on the detection of defects can be noticed. This is reasonable since, as happens in the intermittent mode, when good quality items are still produced there is the possibility that a good quality item is chosen as a sample. Hence, the triggering of the WECO rule is delayed. On the other hand, it is possible that, even if produced, good quality items are never chosen as samples; that obviously implies a reduction of the out of control production time and the number of poor quality items until the problem detection. This justifies the similar results obtained for both the defect introduction modes in the machines of the first station (Figure 4.18 and Figure 4.19). The results obtained would suggest that, when defects are introduced intermittently, a delay in the quality failure detection is a more common situation than the eventuality when times to detection do not changed in comparison with the persistent defect introduction mode. Moreover, it has to be noted that the measurements depend on the percentages of good items produced after the machines go out of control. As regards the particular scenarios analysed, it is worth noting that the impact of the different introduction methods though detectable is not absolutely significant.

4.4.2 Station width

The effect of the width of the station where defects are introduced on the detection ability can be investigated using the results obtained. In fact, Scenarios D3, D1 and D5 introduce defects in a permanent fashion in station 3, 1 and 5, respectively; for the same stations, scenarios D2, D4 and D6 simulate the intermittent introduction of defects. The three stations considered differ for the number of operating machines; in particular, station 3 is the narrowest and station 5 the widest (Figure 4.20).

The number of poor quality items produced until the quality failure detection is shown in Figure 4.22. The results are grouped in Figure 4.22 in order to isolate the station width effect from the defect introduction mode effect. The pattern of the 95th

percentile, the mean and the 5th percentile of the distribution of the production time of poor quality items to detection is represented in Figure 4.23.

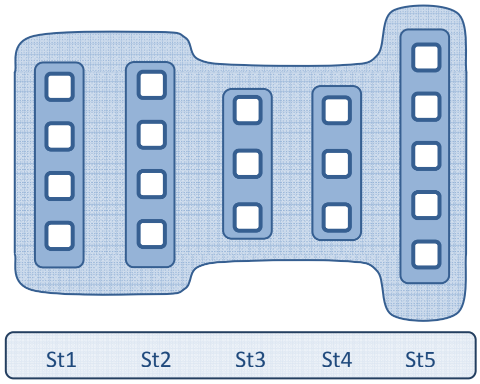


FIGURE 4.21 SYSTEM CONFIGURATION FOR ALL DEFECT INTRODUCTION SCENARIOS.

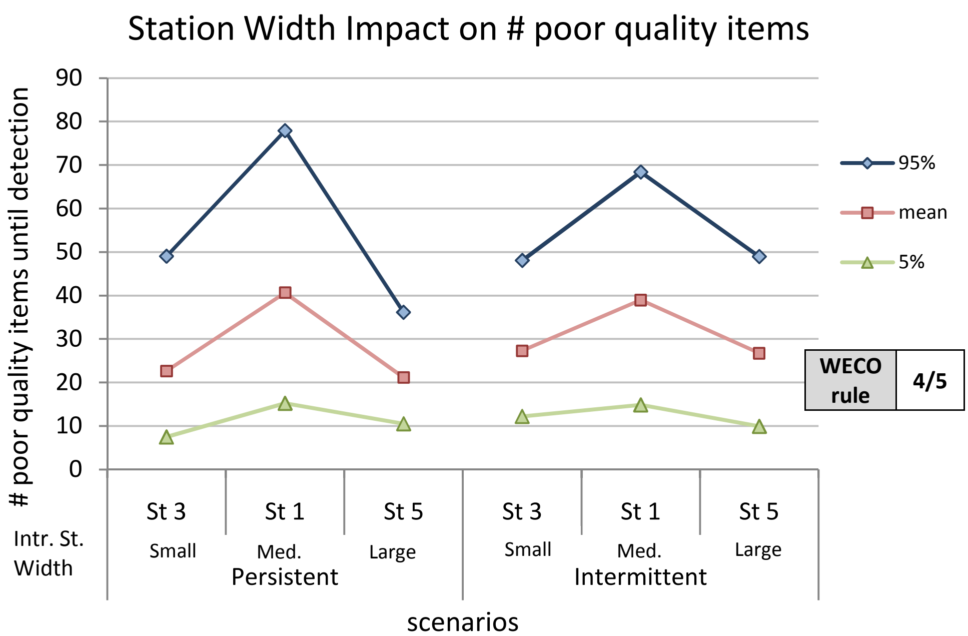


FIGURE 4.22 NUMBER OF POOR QUALITY ITEMS: COMPARISON BETWEEN DEFECT INTRODUCTION IN STATIONS OF DIFFERENT WIDTH.

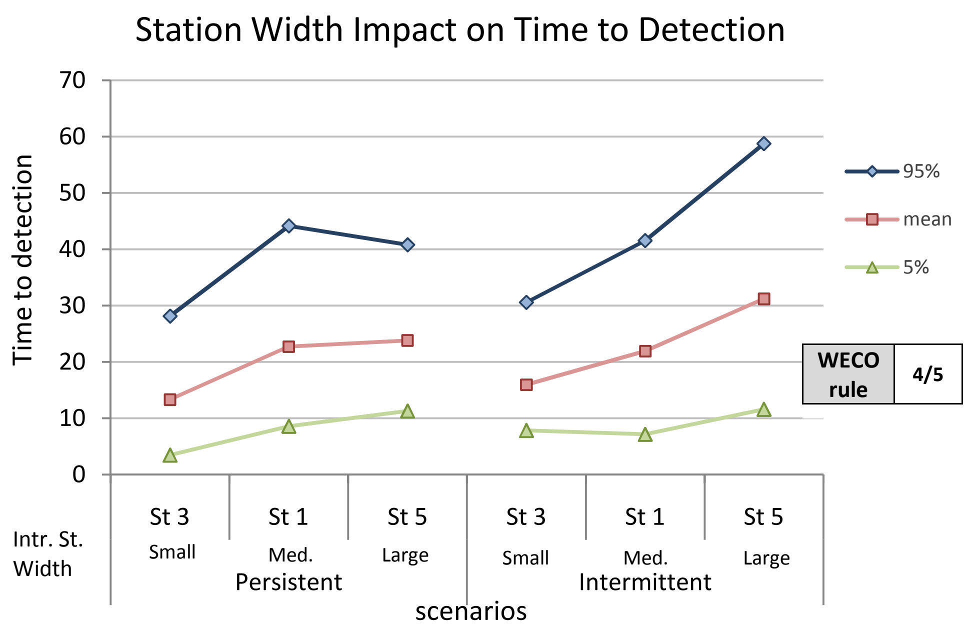


FIGURE 4.23 NORMALIZED TIME TO DETECTION: COMPARISON BETWEEN DEFECT INTRODUCTION IN STATIONS OF DIFFERENT WIDTH.

From the graphs, it is clear that, as regards the average values, the narrower the station is the faster the quality failure detection. In absence of cross flow, this would also mean a fewer the number of poor quality items produced until the machine is shut down. This happens independently of the defect introduction method used. In fact, provided that the workload of each machine in a station is reasonably balanced and the overall number of items processed in each station is approximately the same, the narrower a toolset is the higher the load of each tool. Then, if a machine goes out of control, the overall number of poor quality items produced will increase and in particular the number of poor quality items eligible to be measured at the end of the segment increases. That increases the probability that a poor quality item is sampled. As happened in the previous analysis, a comparison between the number of items produced in different stations is not meaningful due to the different unmonitored flow which crosses the stations.

Simulating the introduction of defects in station 2 would have given the possibility to verify, by means of a comparison with scenario D1 results, that when the number of machines in a station is kept unchanged the performance measures analysed do not vary. However, given the results illustrated in Section 4.3.4, this analysis was considered repetitive and its outcome predictable. For these reasons, further investigation on this issue was not planned.

4.4.3 Effective repair events impact

The introduction of defects in experiments with the inclusion of the PM policy adopted in the factory was simulated on the basis that when routine maintenance events are carried out even if the inspection feedback has not yet pointed to any quality failure of the machine, a certain improvement of the machine functionality is still obtained. In most cases, however, these events do not completely solve the quality problem and the production of poor quality items is still possible. In fact, quite surely, after a short period of an apparent good quality production, defective items will be produced. It is presumable that the repair events which are supposed to improve the behaviour of the machine have higher duration than the usual maintenance events. In this analysis, PM events longer than two hours were assumed to restore an apparently normal functioning of the machine. Scenario D7 was used to investigate the impact of effective PM events in the fifth station of the segment. The results are compared with scenario D5 and D6, which simulate the introduction of defects in a permanent and intermittent fashion in the same station.

Figure 4.24 shows the number of poor quality items produced before the quality failure is detected for these three scenarios. The persistent introduction has the lowest number of poor quality items in comparison with the intermittent introduction and the persistent mode with effective repairing events. The last mode represents the worst behaviour in quality terms since the production of poor quality items is not easily detectable. This is due to the partial re-establishment of the in-control state of the machine so that the consequent production of good quality items for a period delays the detection of the quality failure. At the same time, since the machine is still experiencing a quality failure, poor quality items are still produced and their production persists for longer even if it is spaced out with a relevant production of non-defective items.

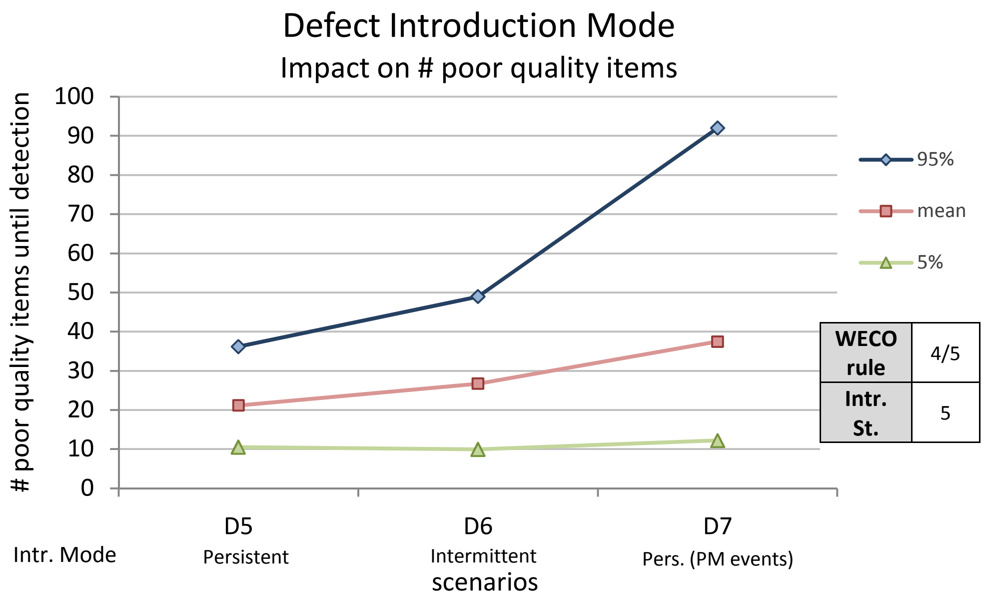


FIGURE 4.24 NUMBER OF POOR QUALITY ITEMS: COMPARISON BETWEEN DIFFERENT DEFECT INTRODUCTION MODES.

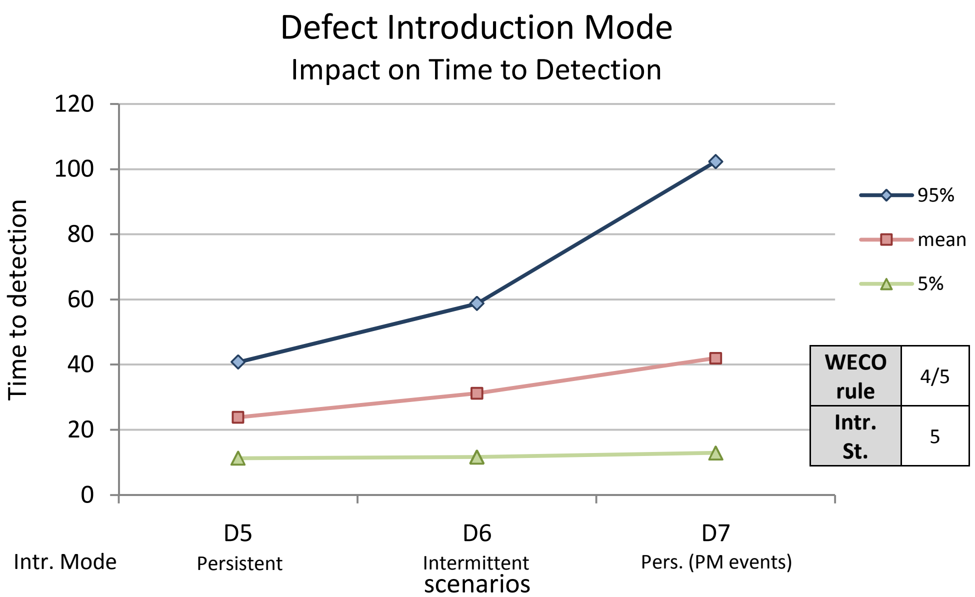


FIGURE 4.25 TIME TO DETECTION: COMPARISON BETWEEN DIFFERENT DEFECT INTRODUCTION MODES.

Paradoxically, a repairing event partially correcting a quality failure reduces the chances to properly restore the machine functioning in a short time and creates an amplification effect on the production of poor quality items. The same pattern characterises the time to detection, as Figure 4.25 shows.

4.4.4 WECO rules impact

The use of stricter rules in quality control policies has two controversial elements. It is generally supposed to catch any quality problem faster but at the same time there is a higher probability that false alarms will cause unnecessary interruption of production [34]. In this study the effectiveness of a stricter rule in detecting quality problems was investigated (Scenario D8). The stricter WECO rule, whose effects are analysed here, would stop the machine when 2 out of 3 items processed by that machine fail the inspection. Figure 4.26 and Figure 4.27 show how dramatic the effect is of using the stricter rule in comparison with the currently used rule. A reduction of almost 75% is obtained for both the number of poor quality items produced until detection and the time to detection.

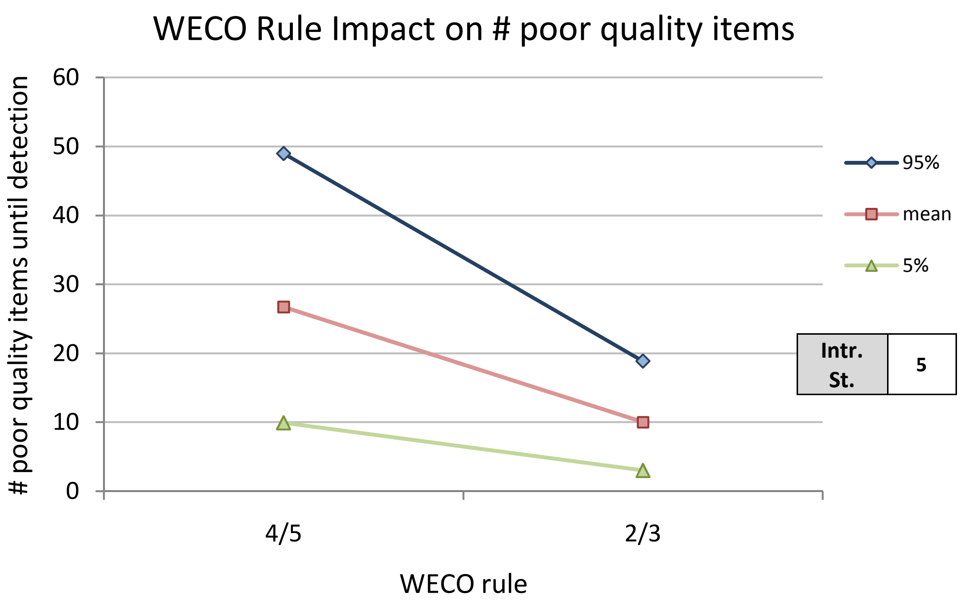


FIGURE 4.26 IMPACT OF A STRICTER WECO RULE ON THE NUMBER OF POOR QUALITY ITEMS BEFORE DETECTION.

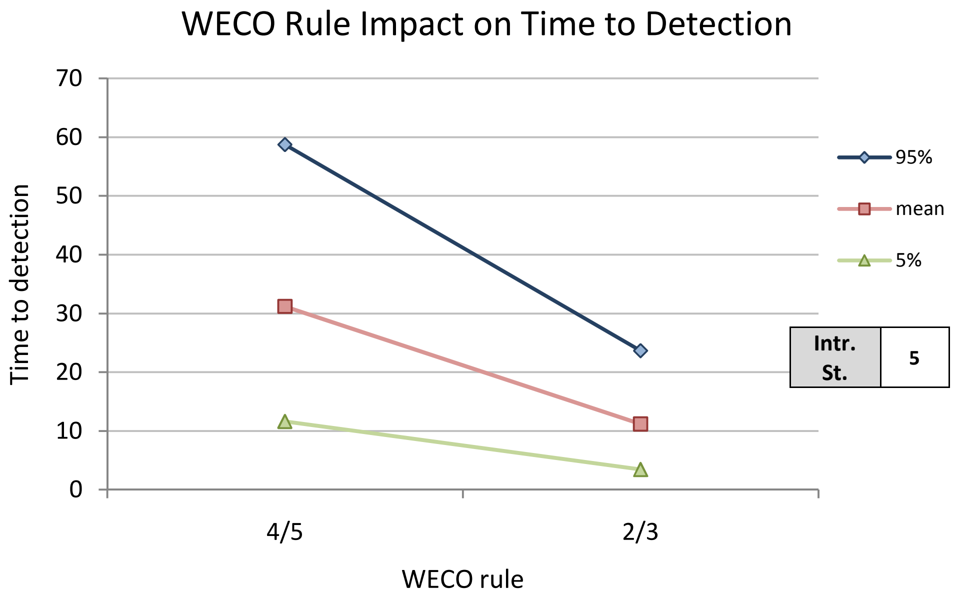


FIGURE 4.27 IMPACT OF A STRICTER WECO RULE ON THE TIME TO DETECTION.

As regards the false alarm analysis, the fact that in these particular scenarios, D6 and D8, the machines are supposed to be interested by quality failures does not allow a proper investigation. However, an idea of the impact of false alarms on the effectiveness of the decision rules of a sampling policy is given in the following section where the introduction of Type II errors is considered.

4.4.5 Type I and II errors

The introduction of poor quality items while a machine works properly was simulated to investigate the impact of the Type I error on the efficacy of both the sampling policy and the decision rule adopted. The defectives were randomly produced based on a given probability, which in this case was 5%. The fifth station was chosen for the introduction of poor quality items in the segment. The results obtained for scenario D6 were modified for this analysis with the help of a random number generator through which the random assignment of an item quality status was performed (Figure 4.28).

Figure 4.29 shows the histogram of the poor quality items produced in timeframes of 750 hours. The histogram was generated by the results obtained from each of the 100 replications available for this scenario. Figure 4.30 reports the histogram of the number of the poor quality items chosen as samples in the same timeframes. In spite of the fact that some poor quality items were produced and sometimes even measured, as Figure 4.30 reveals, the WECO rule adopted, which is 4 defectives out of 5 successive samples, was never satisfied. That gives confidence in the effectiveness of this WECO rule against false alarm signals.

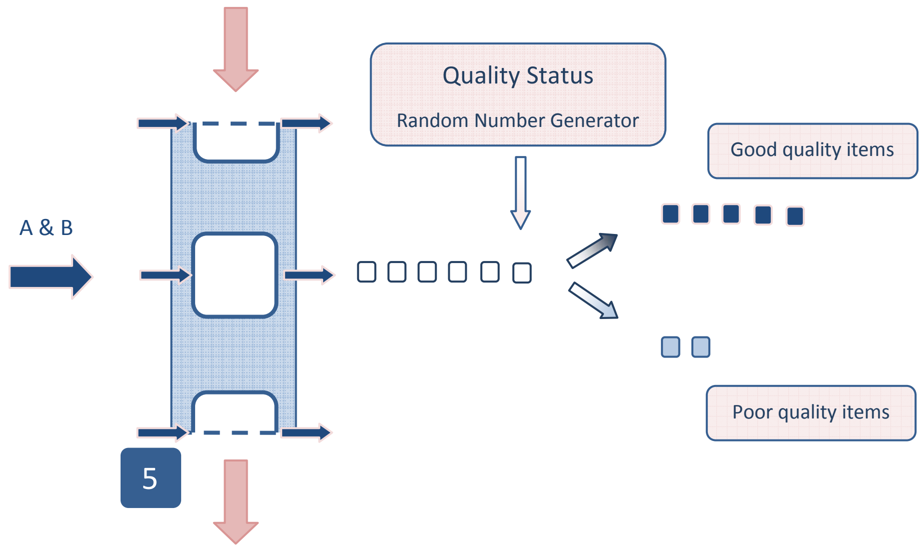


FIGURE 4.28 ASSIGNMENT OF QUALITY STATUS BY MEANS OF AN EXTERNAL RANDOM NUMBER GENERATOR.

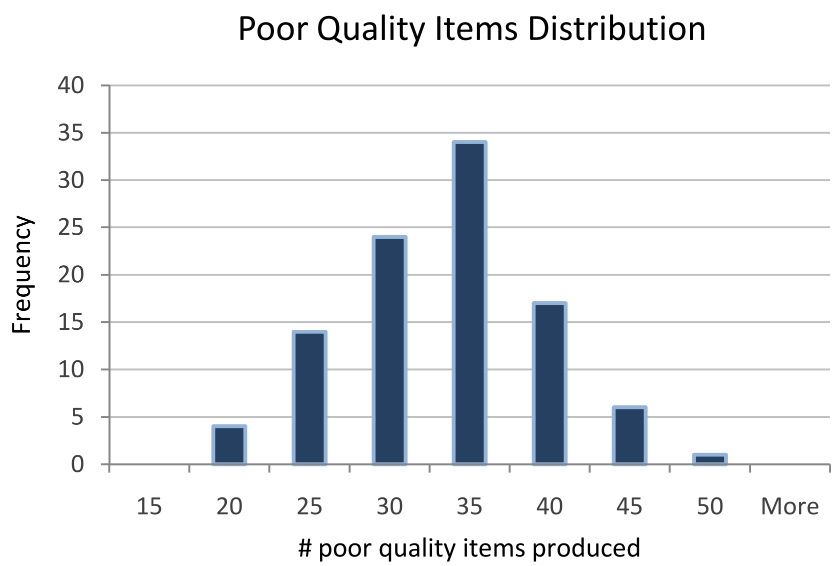


FIGURE 4.29 HISTOGRAM OF THE NUMBER OF POOR QUALITY ITEMS PRODUCED IN 750 HRS

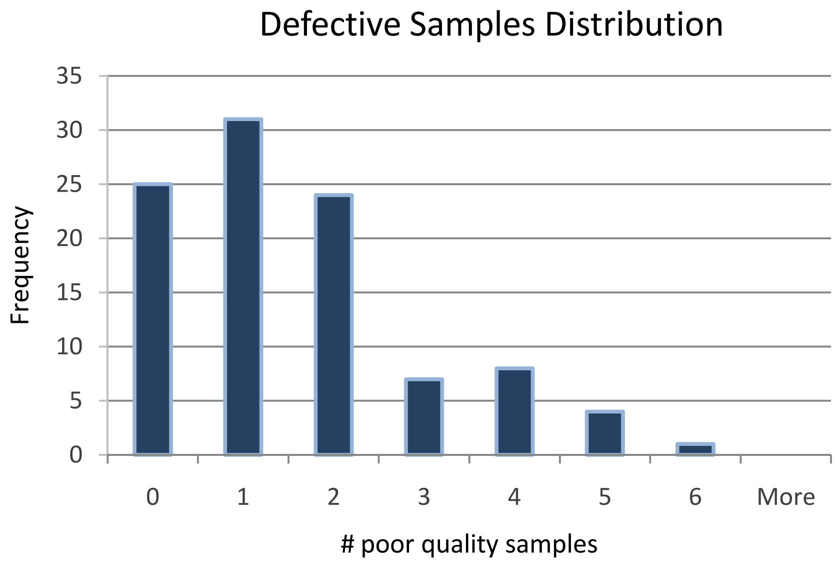


FIGURE 4.30 HISTOGRAM OF THE NUMBER OF POOR QUALITY ITEMS MEASURED IN 750 HRS.

The same results were used to test the effectiveness of the 2 defectives out of 3 consecutive sample rule. The better efficiency of this rule in detecting quality failures presents some drawbacks when the Type I error is considered. In fact, in 15 cases the rule is triggered, hence, the machine would have been shut down if that rule had been implemented in the sampling strategy. The 15 cases correspond with 0.7% of the rule implementations in the total production time simulated. The number of rule implementations is equal to the number of samples (2135 for this particular scenario). The percentage of times for which the WECO rule was triggered, 0.7%, is easily predictable, since the quality status of a sample can be considered a Bernoulli random variable. In fact, since the rule has been applied *a posteriori,* the Bernoulli theorem suggests that the probability of having the rule satisfied is equal to

$$P(x \geq 2) = \sum_{k=2}^{3} \binom{3}{2} 0.05^{k} 0.95^{3-k} = 0.725\% \qquad \textbf{(4.1)}$$

The frequency of false alarm recurring with this rule is quite relevant, considering that the poor quality items constitute only 5% of the whole production. However, it still proves better than the scenario which would be generated by the adoption of the strictest possible shut down decision rule, which is the machine operating interruption as soon as a defective item is detected. In this case, the results suggest that 5.4% of the samples would fail the inspection and would prevent the machine from maintaining production under controlled conditions. The last result is quite reasonable, since in a random sampling policy, the percentage of the poor quality samples should reflect the percentage of poor quality items produced, which is exactly 5% in this case.

Along with Type I errors, Type II errors, usually indicated with β, can be experienced while sampling. They consist of sampling good quality items coming from an out of control machine with the consequent wrong conclusion that the process is in-control. That is possible for the natural dispersion of the production process and represents the opposite situation as the one beforehand analysed. Sometimes measurements errors caused by a relatively low reliability of an inspection machine are still referred to as Type II errors.

The presence of Type II errors ranging in the interval 1-5% (β values) have been considered during the post-processing of scenario D1. Relative to the original decision rule adopted and as expected from the previous results, this error does not impact the measurements at all; therefore, it does not cause any delay in the triggering of the shut down event. Further investigations on higher β values were considered inappropriate, since it would not be realistic in a high quality standard manufacturing environment. The results obtained look reasonable since according to the Bernoulli distribution the chance of having four poor quality items out of five measured items is less than 3*10-3%, when β=0.05. However, it is not so reasonable that no poor quality items were reported.

Table 4.4 summarises the main results found in this section.

TABLE 4.4 MAIN FINDINGS OF SECTION 4.4.

Parameter	Impact on Quality Failure Detection
Intermittent Defect Introduction	Generally, the production of good quality items while the machine experiences a quality failure delays the detection of the quality issue. The number of poor quality items might reduce with respect to the case of persistent production of defectives.
Station Width	For a narrow station, a quality failure is detected faster.
PM policy	PM events, when not able to reveal the quality failure, delay the detection of the issue.
WECO Rules	Dramatic reduction of both the # poor quality items produced before detection and the time to detection when the 2 out of 3 sample rule is adopted with respect to the 4 out of 5 rule. However, the probability of false alarms arises (from $3*10^{-3}$% to 0.7%).

4.5 Conclusions

The analysis conducted in this chapter by means of a simulation approach provided interesting and sometimes apparently counterintuitive results relative to the impact of some control parameters on the monitoring capabilities of the sampling strategy under investigation. The availability of a reliable simulation model permitted the investigation of scenarios which would have been difficult to realise in the real factory. It also proved useful for reducing the time needed to obtain statistically valid results for those scenarios which could be more easily implemented in the real manufacturing environment, particularly when the distribution is more significant than the mean.

A very interesting finding regards the negative effect of the line speed on the monitoring frequency of the machines in the segment. In fact, when the line speed is reduced by reducing the production volumes that cross the segment, the time between samples increases as a consequence of the reduced inter-arrival times. In the same circumstances, the number of unsampled items between consecutive samples is not affected by the line speed. The number of parallel machines in a station plays an interesting role on the time between samples in the moment the different production capacity determines variations of the inter-arrival times at a machine level. That occurs when the inter-arrival times at a station level are kept unchanged while the number of available machines is varied. Another relevant result found is the fact that the reduction of the sampling intervals seems to be the only way to relevantly reduce the number of unsampled items between consecutive samples. In particular, the number of unsampled items proves quite sensitive to the variations of the sampled volume fraction. This means that the presence of an unmonitored flow in the station magnifies the effects of long times between samples in terms of items produced.

The analysis of the defect introduction modes revealed that a persistent production of defectives helps to reduce the time needed to detect a machine quality failure; the eventual production of good quality items while the machine is out of control could cause delays in the detection of the problem. It was found that the worst case scenario from a quality control viewpoint occurs when a machine is partially fixed by ordinary maintenance operations which don't address the failure but partially restores a temporarily good functioning of the machine. The same parameters investigated in the previous analysis were also studied in terms of the impact they have on the defect detectability and results compatible with the results previously obtained were found.

The sampling strategy was also assessed with respect to the WECO rules implemented. The WECO rule originally adopted in the factory proved quite effective in avoiding false alarms but not as efficient as stricter rules for quality failure detection. The machine shut down when two out of three consecutive samples fail the inspection seems to be a better compromise in terms of monitoring performances and false alarm dangers.

Type I and II errors were also investigated. As regard the Type I error, it was interesting to find that the results obtained by means of the simulation approach are well predicted using some statistical considerations. In fact, the binomial distribution can provide the same information revealed by the simulation results about the impact of the Type I error on the actual danger of sampling poor quality items while the machine is under control, when different WECO rules are adopted. Finally, the Type II error does not seem to create any inconvenience relatively to the case scenarios analysed.

The individuation of the parameters that have the greatest influence on the quality risk related performance measures and the understanding of their impact on these measures represent a relevant result obtained from the analysis of the simulation output. Even though a quantitative analysis has been presented, the relationship between the control parameters investigated and the performance measures has been essentially explored in a qualitative fashion; in fact, the impact of a particular combination of parameters can be roughly foreshadowed considering the effect that every single parameter has on the time and the number of items between consecutive samples. This prediction might prove poor since based on partial perspectives; the interaction between parameter variations is difficult to assess, above all, when they cause opposite effects on the final measures. This is because an understanding of the impact of the single parameters is not supported by a quantitative estimation that could reveal which of the different effects eventually prevails. A quantitative assessment requires the simulation of the particular scenario to be investigated and the comparison between different parameter settings can prove a time consuming process. The availability of analytical models able to translate the qualitative relationships analysed into quantitative expressions would be extremely convenient for both speeding up the analysis process and pose the basis for the development of standard procedures for keeping the quality risk under the desired control level.

The next chapter reports the investigation of the simulation results from a mathematical perspective and the derivation of analytical models for the quantitative prediction of the quality risk associated with a sampling strategy.

Chapter V

Quality Risk Prediction: an Analytical Approach

5.1 Introduction

As the analyses conducted in the previous chapter show, the simulation model of the production segment under investigation proved to be a fundamental tool for gaining insight into the system behaviour and exploring scenarios which would have been difficult to implement in the real system without causing relevant inconveniences. However, even if its availability represented a great advantage for evaluating the impact of some control parameters on the performances of the sampling strategy adopted, it is also true that the prediction of the quality risk using such a tool is inefficient. In fact, statistically valid results require several simulation runs and data processing by means of different software; in other words they are not immediately available and the time needed to obtain them could be greater than the time available to make decisions. Moreover, due to its rigidity, the cost of ownership of the model could prove quite high. For these reasons, the opportunity to derive from the simulation results analytical models for the prediction of the performance measures previously considered was explored. To this end, a brief literature survey is presented in the next section. This survey highlights merits and limits of the analytical approaches. It also suggests that the fusion of simulation and analytical approaches has proved to be particularly effective in different research fields.

5.1.1 Methodology considerations

The advantages and limits of the analytical approach are basically complementary with respect to the limits and advantages of the simulation approach. That means most of them have already been illustrated from the simulation world viewpoint (Section 3.2.1.1). A few more are presented in this section.

The fact that, whenever applicable, analytical models are generally considered preferable to simulation approaches [139] could be a signal that these models are easier to use and usually provide more immediate answers. In fact, apart from the cases when numerical approaches are needed, the application of the analytical procedures is cheaper in comparison to the solution costs required by a simulation approach [140]. Here the concept of costs is generalised to include time, resources and skills necessary to obtain solutions. Even building the model itself and developing solution procedures can have, on average, lower costs for an analytical approach than for simulation [139].

However, there is a cost to pay for these merits. The level of details that is possible to include in the analytical models is limited by the feasibility of the solution procedure. The assumptions made sometimes defy reality and, as a consequence, the resulting model, even though alluring and perfect from a theoretical point of view, is barely useful in real life; in the end, the model developed does not reproduce the real behaviour of the system.

Conversely, the process of abstraction from reality sometimes does not compromise the efficacy of the model in gaining insight into the system dynamics. In fact, if the simplifying assumptions regard marginal aspects of the problem or consist of generalisations whose impact on the solution can be predicted and eventually corrected, sufficient model realism and enough accuracy for the desirable performance measures can be obtained. The validation of these simplified analytical models should be carried out against the real system, or against its faithful representation. That is the idea based on which Ignall et al. [139] suggest that the use of simulations to test other mathematical models can be considered conceptually analogous to the use of experiments performed in a real system, provided that the simulation model has been previously validated. To support this suggestion, they report four case studies where simulations were used to develop and validate simplified analytical models. Ensuring that an analytical model works as well as a simulation model will result in immediate savings in time and money. Among the other advantages, the authors also consider that an analytical model can be more easily embedded in other models. The methodology proposed by Ignall et al. [139] has been adopted in this study where simulations constitute the basis on which the analytical models have been developed and successively validated.

TABLE 5.1 CLASSIFICATION OF THE SIMULATION/ANALYTIC MODELS IN [140].

Class	Definition
I	*"A model whose behaviour over time is obtained by alternating between using independent simulations and analytical models"* The simulation and the analytical part of the model do not interact during the solution procedure
II	*"A model in which a simulation model and an analytical model operate in parallel over time with interaction during the solution procedure"*
III	*"A model in which a simulation model is used in a subordinate way for an analytic model of the total system"*
IV	*"A model in which a simulation model is used as an overall model of the total system, and it requires values from the solution procedure of an analytic model representing a portion of the system for some or all of its input parameters"*

The combined use of simulation and analytical models can result in great advantages since the limits of one can be overcome by the merits of the other, all resulting in a reduction of costs. Hybrid simulation/analytical models and modelling are investigated in a broader sense by Shanthikumar and Sargent [140] who classify the different possible ways to integrate these two modelling approaches. Table 5.1 reports the definitions proposed in [140] for the four classes identified by the authors. Nyhuis et al. [141] show how the use of a hybrid approach for the prediction of operating curves of different logistic performance measures can be an interesting alternative to the use of simulation. Since analytical and numerical methods are not available, Wang [142] optimises a static and dynamic model for the definition of the condition monitoring interval by means of a hybrid approach. Byrne and Bakir [143] illustrate the benefits of the hybrid approach in comparison to using either simulation or analytic methods alone for the multi-period, multi-product, production planning problem. Improvements of their approach are also available in [144] and [145].

The analysis conducted in this chapter has been fundamentally based on the intention to embed the simulation results in the shape of analytical models. The approach intended to be used is similar to the first class of hybrid approaches defined in [140], since apart from the initial results made available from the simulation runs, no interactions is meant to happen between the simulation and the analytical models during the solution procedure.

5.1.2 Objectives

The models developed in this chapter focus on two different aspects. First, attention was paid to the prediction of both the time between samples and the number of unsampled items between two consecutive samples in terms of average values. The formulae obtained provide useful insights on the average risk of not monitoring each of the machines operating in the segment. However, they are not able to quantify in terms of either time or number of items the quality risk exposure associable with any confidence level. This type of evaluation is only possible when the distribution of the performance measures is available. In fact, the confidence level can be seen as a cumulative probability and, for instance, the maximum number of consecutive items exposed to the risk of not being sampled at that confidence level would be the value corresponding to that cumulative probability. This value is easily derived from the distribution. Hence, the second group of analytical models attempts to predict the distributions. Only the number of unsampled items between consecutive samples was considered in the second part of the study. The reason for this choice was based on the consideration that the number of items between samples was more frequently used as a risk measure by the production staff in the organisation supporting this research. Moreover, given the close relationship between the two performance measures, as was proved by the formulas for the average values, it originally made sense to focus the analysis on one of the two adopted measures. Finally, another reason for the choice came from the fact that the number of items is a discrete measure and, hence, theoretically, it should be simpler to deal with it rather than with the time between samples, which is obviously a continuous random variable.

The analysis relative to the prediction formula for the average values of the time between samples and the number of consecutive unsampled items will be illustrated in Section 5.2. Section 5.41.1 will delineate the development of the prediction models for the distribution of the number of unsampled items between consecutive samples in a non-sampling station. The sampling station case will be analysed in Section 5.5. Finally, considerations about the use of the prediction models will be presented in Section 5.6 for both the evaluation of the quality risk associated with a sampling strategy and the definition of sampling parameters able to satisfy quality risk constraint.

TABLE 5.2 SPECIFICATIONS OF THE NEW SCENARIOS SIMULATED.

<table>
<tr><th rowspan="2">Scen.</th><th rowspan="2">WIP turn</th><th colspan="6">Number of Machines</th><th colspan="2">Product Sampling Intervals</th><th colspan="2">Inter-Arrival Rate</th><th rowspan="2">Global Sample Rate</th></tr>
<tr><th>St 1</th><th>St 2</th><th>St 3</th><th>St 4</th><th>St 5</th><th>Insp.</th><th>A</th><th>B</th><th>A</th><th>B</th></tr>
<tr><td>1</td><td rowspan="2">Low</td><td rowspan="4">4</td><td rowspan="4">4</td><td rowspan="4">3</td><td rowspan="4">3</td><td rowspan="4">5</td><td rowspan="4">4</td><td>Large</td><td>Large</td><td>High</td><td>High</td><td>Low</td></tr>
<tr><td>2</td><td>Small</td><td>Small</td><td>High</td><td>High</td><td>High</td></tr>
<tr><td>3</td><td rowspan="2">High</td><td>Large</td><td>Large</td><td>Low</td><td>Low</td><td>Very Low</td></tr>
<tr><td>4</td><td>Small</td><td>Small</td><td>Low</td><td>Low</td><td>Low</td></tr>
<tr><td>5</td><td rowspan="2">Low</td><td rowspan="4">5</td><td rowspan="4">5</td><td rowspan="4">3</td><td rowspan="4">4</td><td rowspan="4">6</td><td rowspan="4">5</td><td>Large</td><td>Large</td><td>High</td><td>High</td><td>Low</td></tr>
<tr><td>6</td><td>Small</td><td>Small</td><td>High</td><td>High</td><td>High</td></tr>
<tr><td>7</td><td rowspan="2">High</td><td>Large</td><td>Large</td><td>High</td><td>High</td><td>Low</td></tr>
<tr><td>8</td><td>Small</td><td>Small</td><td>High</td><td>High</td><td>High</td></tr>
<tr><td>9</td><td rowspan="2">Low</td><td>5</td><td>5</td><td>3</td><td>4</td><td>5</td><td>5</td><td>Large</td><td>Large</td><td>High</td><td>High</td><td>Low</td></tr>
<tr><td>10</td><td>5</td><td>5</td><td>3</td><td>4</td><td>7</td><td>5</td><td>Large</td><td>Large</td><td>High</td><td>High</td><td>Low</td></tr>
<tr><td>11</td><td>Low</td><td rowspan="2">5</td><td rowspan="2">5</td><td rowspan="2">3</td><td rowspan="2">4</td><td rowspan="2">6</td><td rowspan="2">5</td><td rowspan="2">Very Small</td><td rowspan="2">Small</td><td rowspan="2">High</td><td rowspan="2">High</td><td rowspan="2">Very High</td></tr>
<tr><td>12</td><td>High</td></tr>
</table>

5.2 Prediction of average values

The results available from the ten scenarios simulated for the initial analysis provided a good base on which an attentive analysis of the average values could be carried out. Two more scenarios, scenario 11 and 12, were added in order to have a wider range of values for the global sampling rate. Apart from this rate, which was higher for the new scenarios, scenarios 11 and 12 are similar to scenarios 6 and 8, respectively, as their main characteristics reported in Table 5.2 suggest.

The aim of the analysis was to find a regular pattern in the data which could suggest the nature of the mathematical relationship between the average time between samples and the control parameters previously explored, which basically are the sampling intervals, the station width and the line speed. In fact, due to some considerations made in the previous chapter, more than the line speed, the inter-arrival time at a machine level was considered potentially more likely to have an important affect on the performance measures.

5.2.1 Time between samples

The time between samples was considered first since a greater number of results were available for it. That is due to the fact that the time, unlike the number of items between samples, is not affected by the presence of a cross flow; hence, the results obtained in any station could be useful.

After considering different variables with respect to which the average values of the time between samples obtained from the simulation results could be plotted, the most effective display proves to be a 3D graph that has as the two independent variables the number of samples per unit time and the number of machines in a station (Figure 5.1). The third dimension of the graph, that is the dependent variable, is obviously given by the average time between samples. The effectiveness of the different independent variables was assessed based on the regularity of the resulting graphical representation of the average time between samples.

The twelve scenarios (Table 5.2) used to derive the prediction formula for the average time between samples provided more than twelve points to be plotted on the graph. This is because, for each scenario, results corresponding with different numbers of machines are available. On the contrary, there exists a biunique correspondence between the scenarios and the number of samples per unit time. In order to keep the plot as clear as possible, when the same combination of number of machines and number of samples per unit time was available different times either within the same scenario or in different ones, the average of the corresponding time between samples was plotted in Figure 5.1. This does not affect the quality of the plot since the same combination of parameters provides very similar results independently of the scenario they come from. For example, the results available for the combination of five machines and low line sample rate are shown in Table 5.3; this particular combination of variables has been chosen as an example since it presents the greatest number of results associated with it.

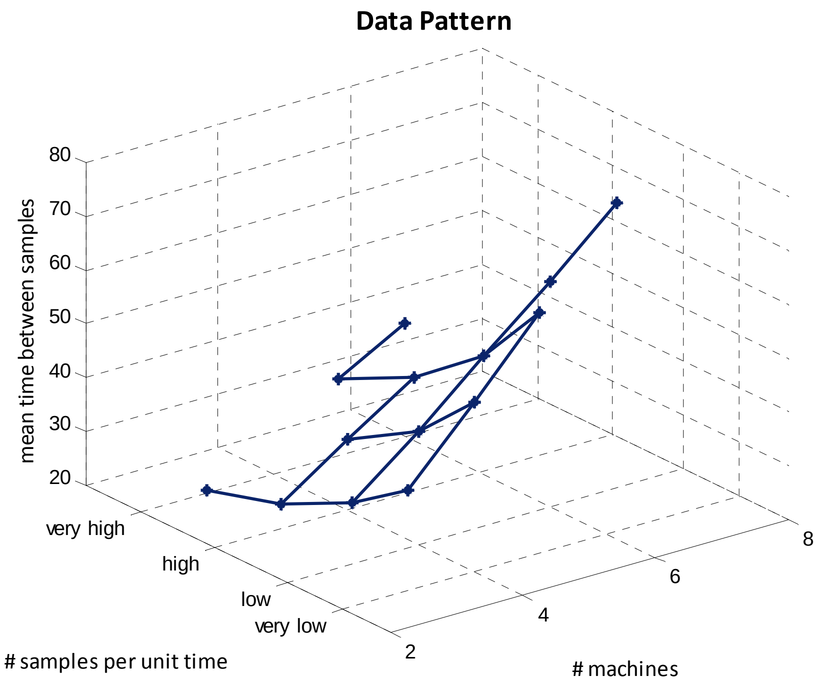


FIGURE 5.1 TIME BETWEEN SAMPLES VS NUMBER OF MACHINES AND NUMBER OF SAMPLES PER UNIT TIME.

TABLE 5.3 MEAN TIME BETWEEN SAMPLES OBTAINED FOR THE COMBINATION OF 5 MACHINES AND A LOW LINE SAMPLE RATE.

Scenario	Station	Time between samples (mean)
1	5	51.64
4	5	51.34
5	1	51.48
	2	51.46
7	1	51.6
	2	51.6

Scenario	Station	Time between samples (mean)
9	1	51.5
	2	51.56
	5	51.52
10	1	51.68
	2	51.76

In order to improve the data representation, the values reported in Figure 5.1 and Table 5.3 were multiplied by a factor different from the one used for the results shown in Chapter 4.

A better understanding of the nature of the patterns was found by projecting the points onto the vertical planes. Relatively simple relationships with each of the parameters emerged from the projected graphs. This provided the basis for the development of the prediction model in that, when the total number of samples per unit time is kept constant, the time between samples for each machine is linear with respect to the number of machines in the station (Figure 5.2). This is a result of the linear relationship between the mean inter-arrival time of items at any one machine and the number of machines in a station for a constant line speed. It is worth remembering that the time between samples is related only to the inter-arrival time of products which are monitored. Keeping the number of machines constant, an inversely proportional relationship between the time between samples and the number of samples per unit time emerged (Figure 5.3). It is intuitive that the greater the number of samples the lower the time between samples; moreover, the nature of the relationship can be explained with the fact that the inverse of the number of samples per unit time represents the time between samples at the station level. Multiplying this value by the number of machines allows the consideration of the delay in the inter-arrival time at a machine level caused by the spread of the overall flow into a station across its machines.

Based on the previous observations the prediction model for the average time between samples, $\Delta T_{samples}$, can be derived very easily. It can be expressed as follows

$$\Delta T_{samples} = \frac{m}{\#\, samples_{unit\ time}} \qquad \textbf{(5.1)}$$

where *m* represents the number of machines in the station and $\#\, samples_{unit\ time}$ the number of samples per unit time, respectively. Since $\#\, samples_{unit\ time}$ is the same for all the stations in the monitored segment, the time between samples varies from station to station depending only on the number of machines.

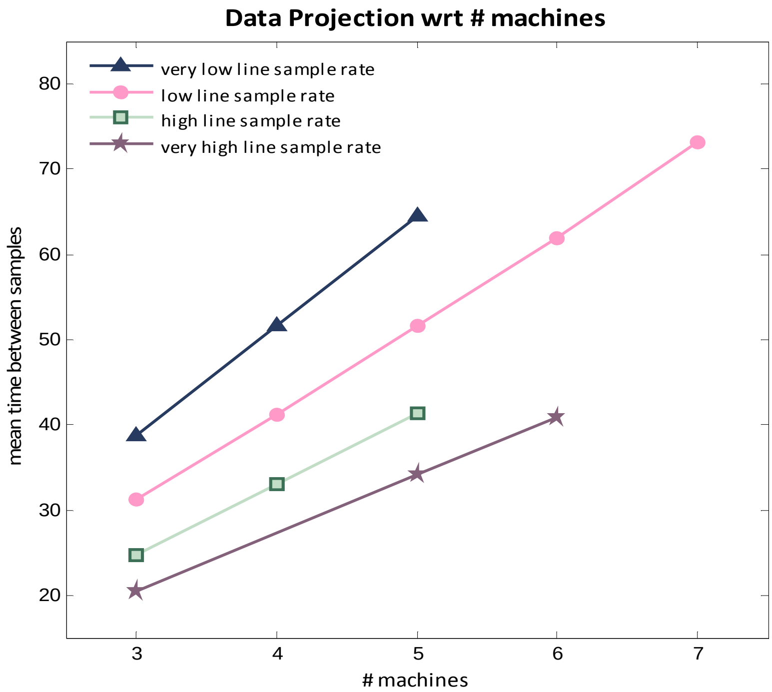


FIGURE 5.2 TIME BETWEEN SAMPLES VS NUMBER OF MACHINES.

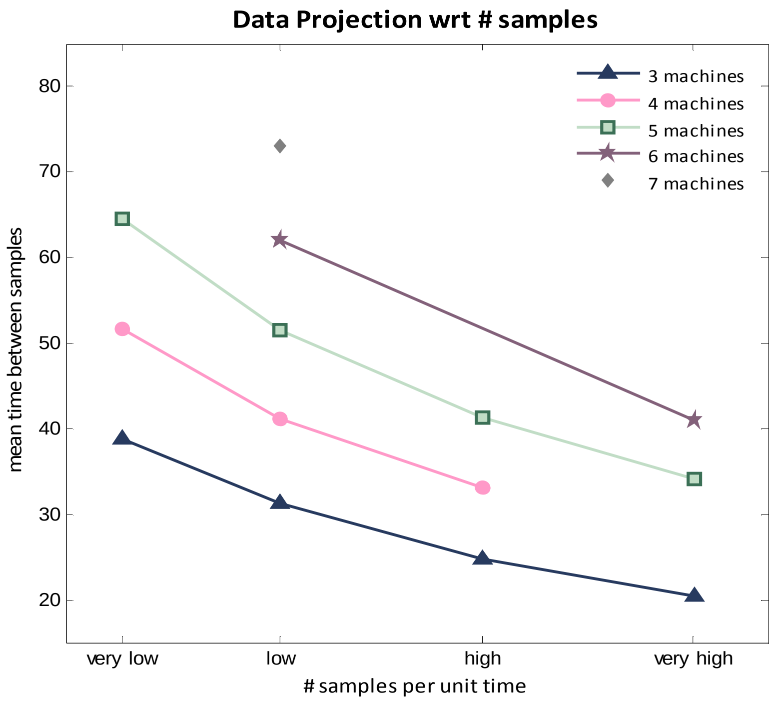


FIGURE 5.3 TIME BETWEEN SAMPLES VS NUMBER OF SAMPLES PER UNIT TIME.

Equation 5.1 is the most compact expression which can be used for the time between samples; however, relating the number of samples per unit time with its affecting parameters, which are the sampling intervals and the inter-arrival times of monitored products, a more basic equation can be found.

$$\Delta T_{samples} = \frac{m}{\sum_i \left(\frac{1}{int.\ arr.\ time_i} * \frac{1}{f_i} \right)} \quad \textbf{(5.2)}$$

where the index *i* refers to all the products undergoing the sampling decision and f_i represents the sampling interval of the i^{th} product, that is the number of items between samples, sample included, in the station where the sampling decision is made.

5.2.1.1 Validation

The formula derived for the time between samples was validated against some simulation results. Other scenarios were simulated so that the validity of the formula could be tested both within and outside the domain on which the prediction model was developed. In particular, the sampling intervals for both the monitored product types were reduced so that a higher number of samples per unit was available; two other sample rates, in between the low and the high number of samples per unit time were also tested. The new scenarios were characterised by a low WIP turn; stations 1, 2, 3 and 5 operated with 5, 4, 3 and 6 machines, respectively.

The surface illustrated in Figure 5.4 is generated by the application of Equation 5.1. The time between samples is represented against the number of samples per unit time and the number of machines in a station. As noted before, this is the easiest way to represent the time between samples; in fact, the use of Equation 5.2 would have been impossible on a 3-D graph, unless at least one of the variables involved would have been kept unchanged. The squares represent the results on which the formula was developed, whereas the circles are the simulation results based on which the validation was conducted. The fact that these points lie on the surface gives confidence in the validity of the prediction model. The relative percentage errors observed for the scenarios analysed are always less than 1.7%. The relative percentage error was calculated as follows:

$$\text{Relative Error } \% = \left|\frac{\text{Predicted value} - \text{Actual value}}{\text{Actual value}}\right| * 100\% \qquad (5.3)$$

where the actual values are the simulation results and the predicted values the results obtained with Equation 5.2.

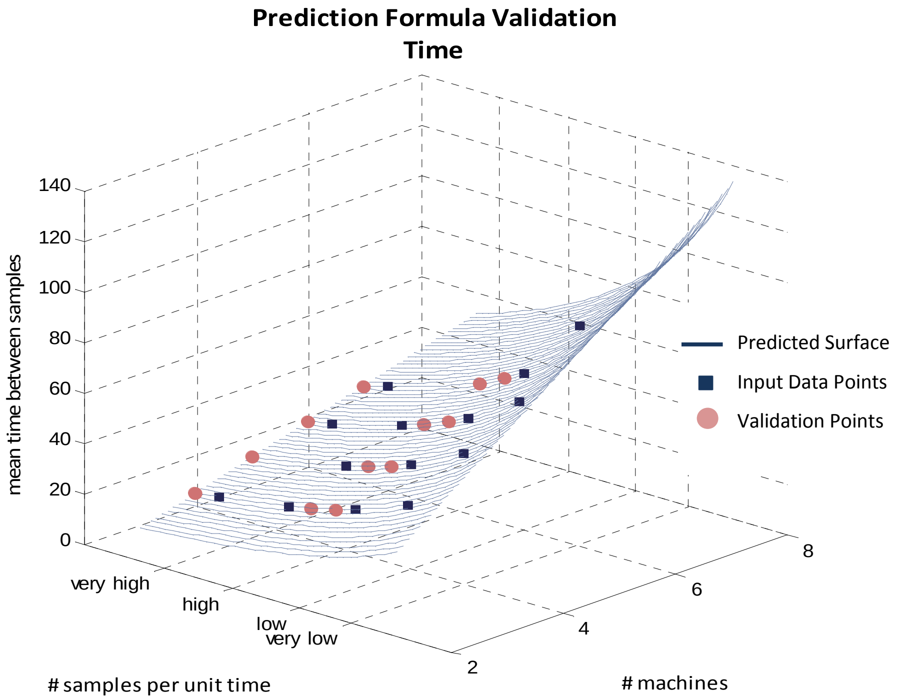


FIGURE 5.4 VALIDATION OF THE FORMULA FOR THE TIME BETWEEN SAMPLES.

5.2.2 Number of consecutive unsampled items

Once the time between samples is defined, the prediction model for the average number of unsampled items can be immediately derived. In fact, the average inter-arrival time at a machine level can be used to convert the time between samples into the number of unsampled items between samples. This obviously works exclusively for the average values. The formula obtained is the following:

$$\begin{aligned}\#\, items_{between\ samples} &= \frac{\Delta T_{samples}}{machine\ \ int.\, arr.\ time_{global}} - 1 \\ &= \frac{\Delta T_{samples}}{m * int.\, arr.\ time_{global}} - 1\end{aligned} \tag{5.4}$$

where the *machine int. arr.* $time_{global}$ represents the overall inter-arrival time for a machine, inclusive of all products under manufacture. The unit decrement in equation 5.4 represents the sampled item. Substituting Equation 5.2 into Equation 5.4 the following expression is obtained:

$$\#\, items_{between\ samples} = \frac{1}{\left[\sum_i \left(\frac{1}{int.\, arr.\, time_i} * \frac{1}{f_i}\right)\right] * int.\, arr.\, time_{global}} - 1 \tag{5.5}$$

Considering the meaning of the denominator, Equation 5.5 can be rewritten in a more compact fashion. Due to the number of parameters involved, the new formulation, reported in Equation 5.6, would prove more convenient than Equation 5.5 when a graphical representation of the pattern of the average number of items between samples with respect to an affecting parameter is needed.

$$\#\, items_{between\ samples} = \frac{1}{Sampled\ \ fraction} - 1 \tag{5.6}$$

Equation 5.5 reveals that the average number of unsampled items is independent of the number of machines in the station. This means that where there is no cross flow to vary the inter-arrival rate between stations, all stations have the same number of unsampled items, which is the average sampling interval in the station where the sampling decision is made minus one.

The presence of the cross flow, which does not affect the time between samples, appears in Equation 5.5, in terms of the *int. arr.* $time_{global}$. The greater this flow, the smaller the global inter-arrival time at the station, hence, the greater the number of items between samples. Therefore, the cross flow acts as a scaling factor; large unmonitored volumes represent a risk in the station since the impact of a quality failure could affect a greater quantity of items.

Equation 5.5 has been specifically developed for the particular sampling strategy described in Section 3.3.2; this means it predicts the average number of unsampled items between samples when the sample size is equal to one. When more than one item is consecutively sampled at the sampling station, that is, when an item is chosen to be sampled the consecutive n-1 items that immediately follow it are also sampled, it is believed that Equation 5.6 can still described the average number of consecutive unsampled items at a non-sampling station. This statement is supported by the results illustrated later on in this chapter (Section 5.4). For the sampling station, a generalisation of Equation 5.5 for this variant of the sampling strategy is not immediate and would require additional experiments. Since the structure of this variant of the sampling strategy substantially differs from the original one, further investigations on the reaction of the measures to variation of the sampling scheme are considered not opportune in this dissertation.

It is worth noting that all the equations developed in this section can be adapted to a batch production provided that the focus is kept on the number of consecutive unsampled batches and information on sampling within a batch is ignored.

Apparently there is no difference in the number of unsampled items between the station with the deterministic sampling and the stations with the random sampling. In fact, there is no element in Equation 5.5 which prevents using it in any situation. As for the average time between samples, the formula was developed by also taking the simulation results obtained in the sampling station into account. Moreover, the nature of the variables involved makes the formulae extendable to more complex combinations of flow; that means, Equations 5.2 and 5.5 should be still valid when more than two monitored product types flow through the segment. This is affirmed based on the consideration that it is the number of samples per unit time that plays the most relevant role in the formulae, no matter how many monitored product types contribute to their determination.

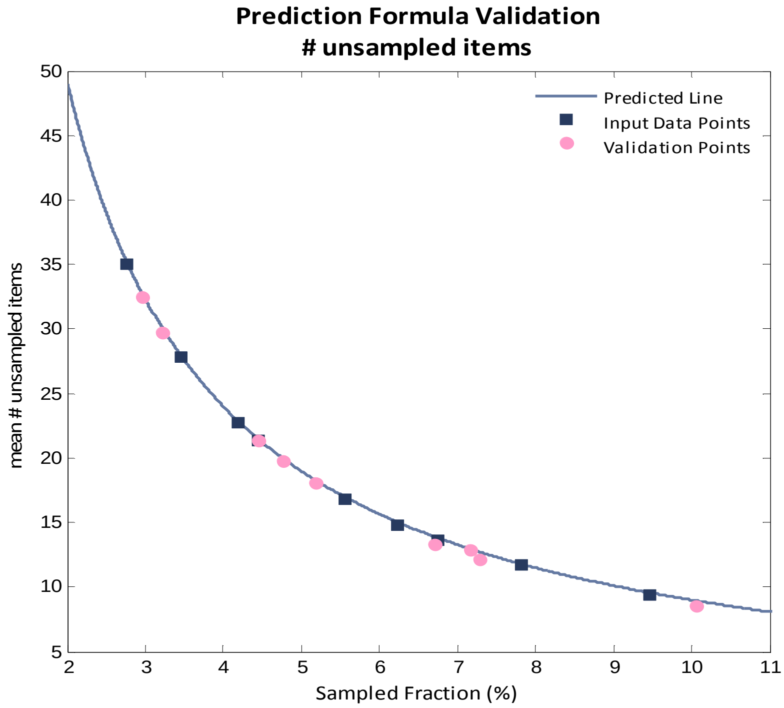


FIGURE 5.5 VALIDATION OF THE FORMULA OF THE AVERAGE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS.

The formula for the average number of unsampled items was validated against the simulation results used for the validation of the prediction formula for the time between samples. As Figure 5.5 shows, a very good fit was found, and for the scenarios investigated the relative percentage error was never higher than 3.5%. Relative to the validation data, the root mean square error was equal to 0.37 [item] and its corresponding relative measure was 2.9%. Along with high prediction accuracy within the domain defined by the observed data, it is interesting to note the capability of the model to predict values outside this domain with the same accuracy level. This gives confidence that the validity of the model is not confined to restricted parameter ranges and can be effectively used to predict the mean number of unsampled items in any production and sampling condition.

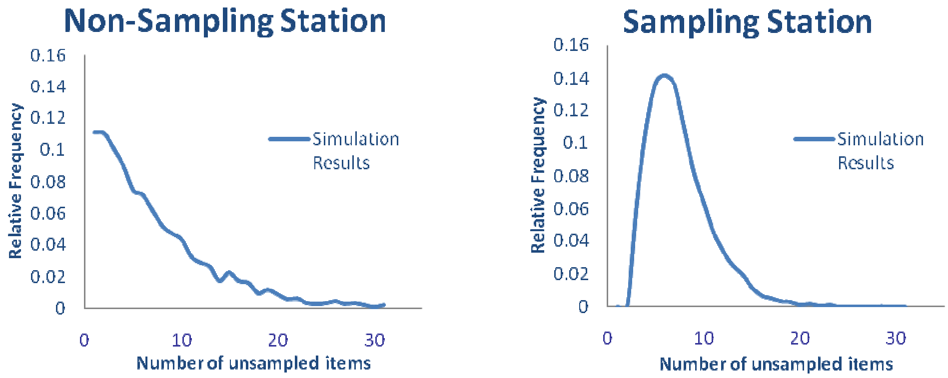
The formulae developed in this section for predicting the average values of the time between samples and the number of items between samples will be referred to as Average Prediction (AP) formulae hereinafter. The approaches based on these formulae will be named algebraic approaches.

Even though the prediction of the average values of both the number of consecutive unsampled items and the time between consecutive samples represents a useful tool for an approximated evaluation of the monitoring effectiveness of the sampling strategy analysed in this research, the quantification of the risk of not monitoring a machine requires a better understanding of the way the number of consecutively unsampled items distribute around the mean value. The rationale behind the development of prediction models for the distribution of the number of unsampled items between consecutive samples under different scenarios is presented in the next sections.

5.3 Distribution of the number of consecutive unsampled items

As noted in the comments pertaining to Equation 5.5, there is no element in the equation which presumes any differences between a sampling and a non-sampling station as regards the number of unsampled items between consecutive samples. However, this is true only when the analysis is limited to the average values. In fact, when the distributions are considered some differences arise. For example, this is intuitive when only one product flows through the segment. In the station where it is sampled, the distribution of the number of unsampled items is obviously degenerative. It consists of only one value, the sampling interval minus one, with probability equal to one. Due to the sequence disorder effect and the multiple stream effect, in all the other stations the number of unsampled items is a random variable and the results obtained suggest that it tends to assume an exponentially shaped distribution. When the product flows get more complicated and, for example, a monitored flow is merged with an unmonitored one, even in the sampling station the regularity of the sampling strategy is lost from a global perspective and an apparent randomness permeates the system.

Figure 5.6 highlights the different shapes of the distribution of the number of unsampled items for a non-sampling and a sampling station. In the case reported, the effect of the cross flows has been ignored so that the scale of the horizontal axes in both the figures could be the same. In this way, besides the different shapes, the different variability of the distribution can also be appreciated. In the sampling station, probably due to the regularity of the sampling plan, the dispersion of the data seems to be smaller than observed for the non-sampling stations.



UTION OF THE
G AND A SAMPLING

Based on these considerations, the analysis of the distribution of the number of consecutive unsampled items was conducted separately for the two classes of stations. In particular, the absence of any deterministic element in the non-sampling stations and, in spite of that, the repetitiveness of the results obtained for the different stations made the resulting analysis of the distribution in these stations paradoxically easier than for the sampling station. As a consequence, the analytical model developed for the non-sampling stations will be shown in the next section. Following that, the analysis conducted for the sampling stations will be presented.

5.4 Non-sampling station case

The relatively regular shape of the distribution of the number of unsampled items between consecutive samples eased the analysis of the non-sampling station case. The exponential shape and the closeness of the mean and the standard deviation allowed the supposition that the exponential distribution could constitute the prediction model. Its parameter, the mean, would be the average sampling interval, which is related to the inverse of the mean number of items between consecutive samples. The only problem is that the exponential distribution is suitable for a continuous random variable, whereas the number of items is obviously discrete. Moreover, the reasons why this distribution could fit the data had to be found in order to justify its applicability and identify the conditions needed for its validity.

Other scenarios were investigated to analyse the impact of the volume of monitored product and the sampling interval on the distributions. The parameter ranges investigated are shown in . A basic model was also developed in order to study the impact of different sources of variability on the shape of the distribution (Section 5.4.4).

TABLE 5.4 PARAMETERS RANGES FOR THE NEW SCENARIOS SIMULATED.

Parameter	Values				
Monitored Volume	20%	40%	60%	80%	100%
Sampling Interval	2	3	4	5	

First, it is worth noting that the cross flow does not change the shape of the distribution for the non-sampling stations. As also Equation 5.5 suggests, its impact is limited to scaling the axis. This is shown in Figure 5.7 where the distribution of the number of unsampled items is drawn for two scenarios differing only by the presence of an unmonitored cross flow. For both the scenarios the sampling intervals of the monitored products is set so that one item out of four is sampled. The distribution relative to the 100% monitored flow scenario proves peakier than the other scenario, which means its mean value and dispersion are less than the other distribution. However, the shape of both the distributions is still decreasing with an exponential pattern. The representation of the distribution of the number of unsampled items, which is discrete, by means of a continuous curve is only due to style preferences. This holds for the rest of this thesis.

Due to the sequence disorder and the multiple stream effects, the order with which monitored items move out from a machine in a station is different with respect to the order with which they exit the machines in the sampling station. Moreover, there is no biunique relationship between the machines of the different stations; this means that even in absence of a sequence disorder effect, the regularity of the sampling plan, even from a single product perspective, would be lost in the stations upstream or downstream of the sampling station. As a consequence, for any of the machines in a non-sampling station, the sampling plan implemented proves random, that is the monitored items are no longer sampled on a regular basis. Hence, the distinction between monitored and unmonitored products does not help the analysis, in the sense that no deterministic pattern which could suggest possible solutions to the problem can be found for any of them. As a result, an unsampled monitored item is not different from an unmonitored item.

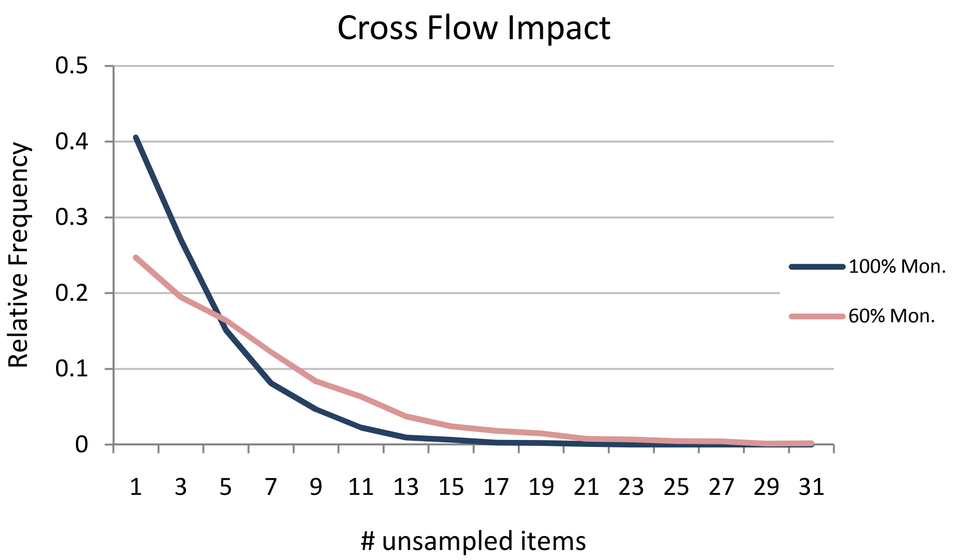


FIGURE 5.7 IMPACT OF THE CROSS FLOW ON THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN SAMPLES.

Another element which can prove useful in the development of the prediction model is the fact that the sequence of products processed by a machine in a non-sampling station is not determined by any logistic rule. This is theoretically justified by the shape of the distributions of the inter-arrival time for the different product types which is exponential. In particular, the memory-less property, which characterises the exponential distribution, makes it possible to state that, since an event is not conditioned by the previous event, the sequence with which items are processed is random. In particular, the sequence of monitored and unmonitored items processed can be considered a geometric process, with the success event being either the first monitored item after consecutive unmonitored items or vice versa. Based on the same reasons, the sequence of sampled and unsampled items is also a geometric process.

The last observation is fundamental. Based on it, it is straightforward to derive the prediction model for the distribution of the number of unsampled items between consecutive samples for the machines of a non-sampling station. In fact, the event of processing either a sampled or an unsampled item can be considered an independent Bernoulli trial which happens with a predefined probability. The probability in question is given by the sampled volume fraction (or the unsampled volume fraction, according to the perspective), which is determined by both the monitored product production volume and the sampling intervals and is constant for each processed items at a non-sampling station machine. If a successful event is intended as the processing of a sampled item and a failure, or an unsuccessful event, as the processing of an unsampled item, the probability corresponding with a given number of unsuccessful events which occur before a sampled item is processed represents the essence of the problem studied in this section. When the number of unsuccessful events is varied, the corresponding probabilities describe a curve which represents the probability mass function of the distribution of the successful event. A basic knowledge of probability theory suggests that the distribution able to describe this particular situation is geometric (See Appendix D). So, the geometric distribution apparently constitutes the prediction model for the distribution of the number of unsampled items in the machines of a non-sampling station, at least in the case when the product flow consists of two monitored products and an unmonitored flow. As the observation just made suggests, the parameter of the geometric distribution, also called proportion, is the sampled fraction of the whole production volume in a station. The proportion, p, in the i^{th} non-sampling station can be calculated using the formula for the average number of unsampled items; it is given by:

$$p_i = \frac{1}{mean\ \#\ unsampled\ items_i + 1} \tag{5.7}$$

It is worth noting that, in general, the inverse of the mean number of observed unsuccessful events (e.g. unsampled items) represents the maximum likelihood estimator of the parameter of the geometric distribution [146].

From Equation 5.7, it is clear that the proportion changes in the different stations due to different unmonitored volumes. Moreover, the fact that Equation 5.5 is generally valid, no matter what the combination of flow is, gives confidence in the possibility to extend the applicability of the geometric model to more complicated combinations of product flows.

As justification for the initial tendency to consider the distribution of the number of unsampled items exponential, it is worth remembering that the geometric distribution is the discrete analogue of the exponential distribution as the two distributions share the memory-less characteristic.

5.4.1 Validation

An immediate test of the validity of the model can be conducted by deriving the distribution mean using the formulation given in Appendix D. The second variant (Equation D.5) has to be chosen, since it expresses the number of events needed on average to observe the first success, which in this case is the sampled items. This variant allows to exclude the sample from the count.

$$\begin{aligned}\mu &= \frac{1-p}{p} = \frac{1}{p} - 1 \\ &= \frac{1}{\frac{1}{mean\ \#\ unsampled\ items} + 1} - 1 \\ &= mean\ \#\ unsampled\ items \end{aligned} \qquad \textbf{(5.8)}$$

The results obtained for the mean of the distribution agree with what was found from the formula of the average number of unsampled items, which reproduced the simulation results.

In order to gain more confidence in the model validity, the predicted distributions were compared with the distributions obtained from the simulation results.

As Figure 5.8 shows, the predicted geometric distribution fits the simulation results very well. The average absolute error for the case shown is as little as 0.21%; whereas, the cumulative absolute error range is 3.27%. The average absolute error was calculated as the arithmetic mean of the absolute errors relative to the first 15 points, being the absolute error intended as the absolute difference between the predicted and the actual value. The cumulative absolute error range represents the difference between the maximum and the minimum prediction errors. The high goodness of fit is not limited to the case shown in Figure 5.8; considering all the scenarios investigated, a very high accuracy is still obtained. It always proves to be higher than 99.6%, in terms of average absolute error (Figure 5.9). The Pearson's chi square test was applied to different scenarios to investigate the statistical significance of the goodness-of-fit for the geometric model. For the scenario illustrated in Figure 5.8, the P-value proved equal to 0.0122, with 27 degrees of freedom characterising the test statistics distribution. The lowest P-value observed for the different scenarios simulated was equal to 0.003 (for 29 degrees of freedom). This result indicates the adequacy of the geometric model for predicting the distribution of the number of consecutive unsampled items in any machine of the non-sampling stations. Since the geometric model is characterised by one parameter, the correction to the degrees of freedom was equal to 2. The number of classes chosen for the test depended on the number of samples available. Whenever more than 5 items were present in a class of width equal to one, the number of classes was increased. It happened that to increase the number of samples in a class, successive classes were grouped together.

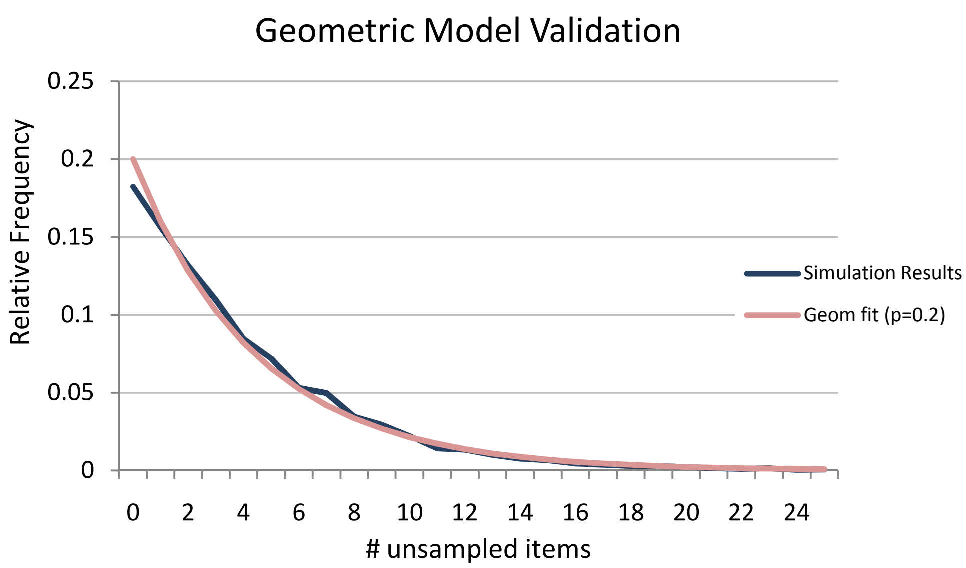


FIGURE 5.8 VALIDATION OF THE GEOMETRIC PREDICTION MODEL FOR THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN CONSECUTIVE SAMPLES.

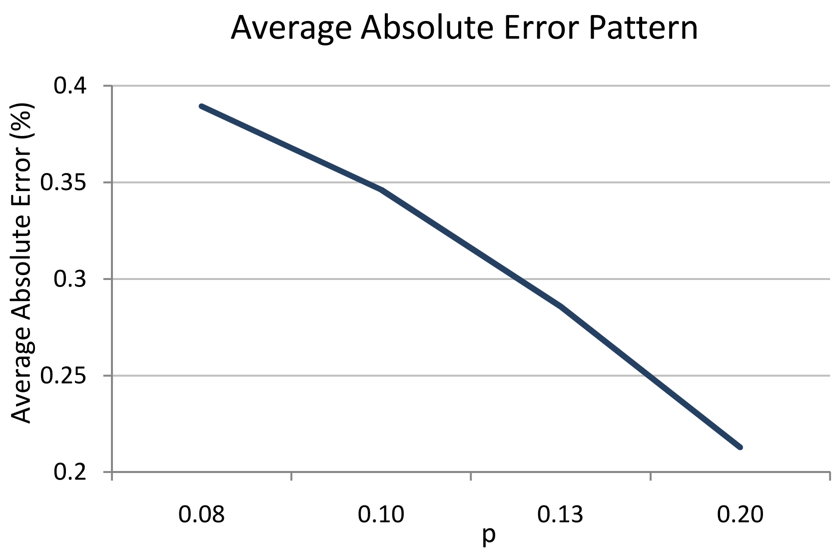


FIGURE 5.9 IMPACT OF THE SAMPLED FRACTION ON THE AVERAGE ABSOLUTE PREDICTION ERROR.

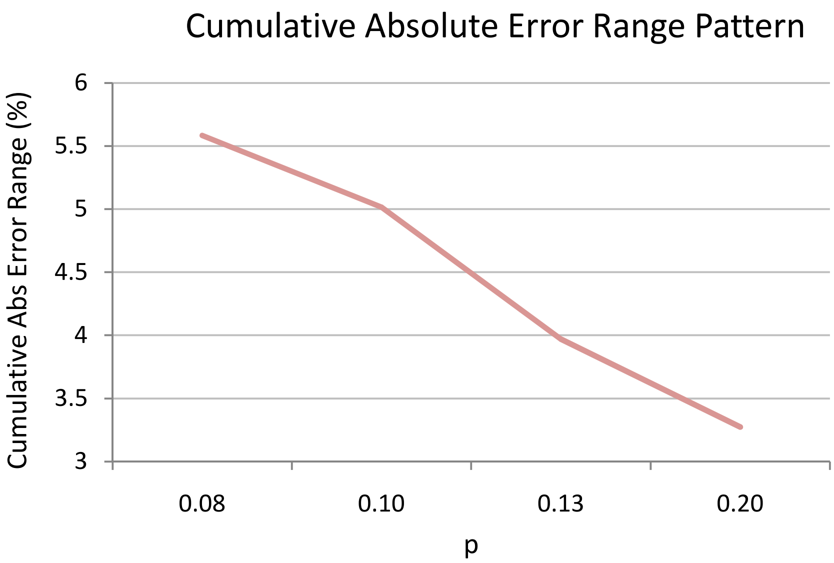


FIGURE 5.10 IMPACT OF THE SAMPLED FRACTION ON THE CUMULATIVE ABSOLUTE PREDICTION ERROR RANGE.

The pattern of the prediction errors was also analysed in order to determine the means by which the magnitude of the proportion impacted the accuracy of the model. For the average absolute error, a decreasing pattern was found with respect to the sample fraction (Figure 5.9). A higher sampled fraction implies a higher probability to process a sampled item and that could reasonably reduce the variability of the number of unsampled items processed between two consecutive samples. The same pattern, with similar slope, was also found for the cumulative absolute error range (Figure 5.10).

5.4.2 Monitored flows merging

Being based on the formula for the average number of unsampled items, which apparently should work for any product flow combination, the geometric model should very likely describe the distribution of the number of unsampled items between consecutive samples for more general product flow scenarios. With the aim of partially proving this, a new scenario was investigated. In particular, the applicability of the prediction model in the case when the monitored flow crossing a station is sampled in two different stations, belonging to different theoretical segments, was analysed. In the simulation model built for this particular analysis, two different products cross the first station. The first product is sampled in the original sampling station, which is the fifth station in the segment. The second product is sampled in a station not belonging to the original segment (Figure 5.11). In order to model the different route followed by the second product, another station was introduced downstream of station 1. Random queuing & transportation, and processing times were considered for this station. One station was considered enough to investigate the effect of merging monitored flows in a station upstream, since the previous analysis showed that the distance from the sampling station has very little impact on the average number of unsampled items and, hence, on its distribution.

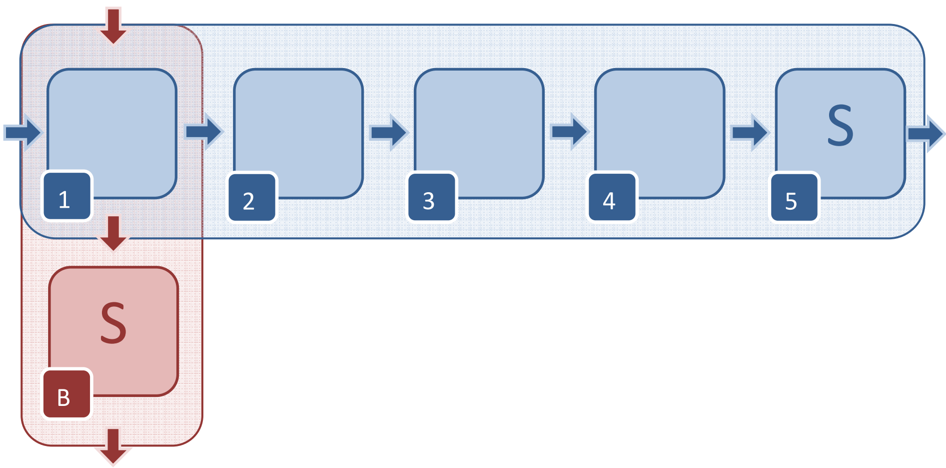


FIGURE 5.11 SCHEMATIC REPRESENTATION OF THE SYSTEM WITH TWO SAMPLING STATIONS GENERATING QUALITY STATUS INFORMATION ABOUT STATION 1.

The two products, with the same inter-arrival times and sampling intervals as used in the modified production model, were also generated in the original simulation model. They were both monitored in the fifth station. The first station was obviously interested by the same sampled fraction as in the previous scenario.

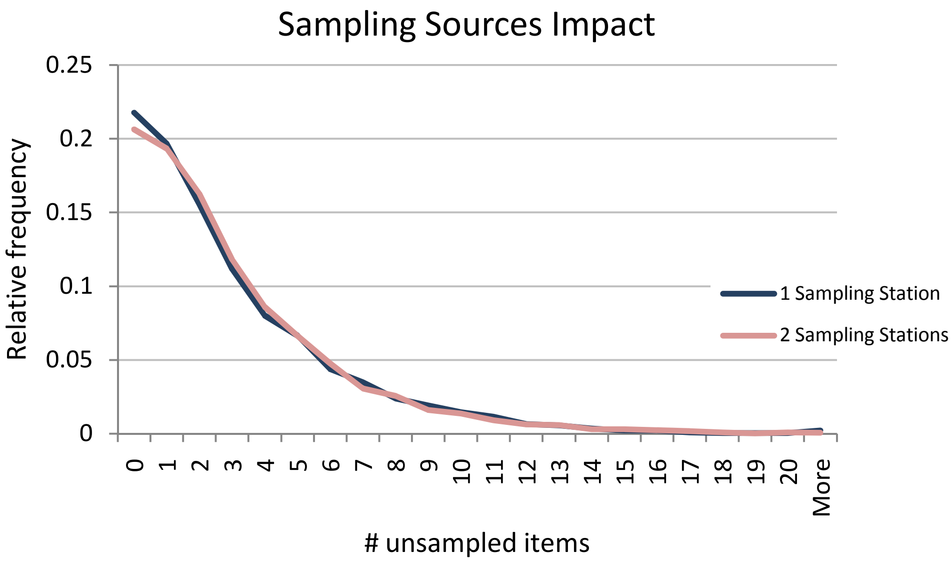


FIGURE 5.12 IMPACT OF DIFFERENT SAMPLING SOURCES ON THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN SAMPLES.

The comparison between the distributions of the number of unsampled items in station 1 for both the scenarios simulated show an extreme closeness of the two curves (Figure 5.12). The small differences between them are very likely a consequence of the different variability which is caused by the different proximity of station 1 from the sampling stations. In fact, whereas the relative difference between the means of the two distributions is only 0.05%, the standard deviations differ relatively by 2.33%. The presence of some items sampled only one station downstream of station 1 slightly reduces the variability of the distribution in comparison with the case when both the monitored products are sampled four stations downstream.

Besides showing that the geometric model is suitable for predicting the distribution of the number of unsampled items in the case when samples are originated by different sampling stations, possibly belonging to different segments, the results obtained also constitute a proof that the formula for the average number of unsampled items (Eq. 5.5) works for more complicated flow combinations as well.

5.4.3 Random serial route impact

The investigation on the validity of the geometric model for the prediction of the distribution of the number of unsampled items between consecutive samples continued with the analysis of another particular case scenario. The case when a station in the segment can be skipped by some serial flow items based on a given probability was considered. This case scenario was inspired by a situation observed in the real system; a station in the real segment processed only half of the serial flow production volume. As already stated in Section 4.2, this led to the choice of simulating two real stations in one station, namely station 4, in the original simulation model and ignoring the results coming from that station due to its particular nature.

The system simulated for this analysis was crossed by two monitored products, A and B. The inter-arrival times of Product B were set to 3.5 times the inter-arrival times for Product A. The second operating station was chosen as the station that could be skipped. The decision whether or not to skip station 2 was made, at a station level, in the moment when the global flow moved out of station 1 (Figure 5.13). Each item, independently of the station 1 machine by which it was processed, had a 50% chance to be routed either to station 2 or directly to station 3. In this case the machine it would be going to visit in station 3 was also randomly assigned.

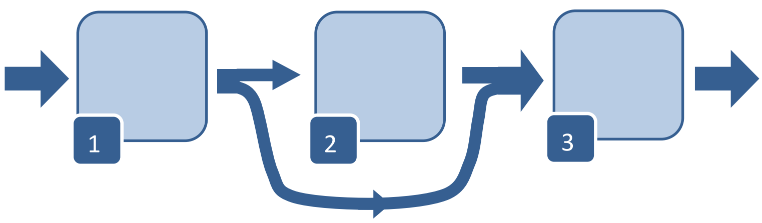


FIGURE 5.13 STATION 2 IS PARTIALLY SKIPPED BY THE SERIAL FLOW.

This scenario was designed so that the impact of random routes, not always strictly serial, followed by the items could have on the shape of the distribution of the number of consecutive unsampled items could be investigated. These random routes obviously create a difference in the number of available samples for the skipped station, so that a variation in the shape of the distribution could be expectable.

The results obtained contraddicted this consideration since irrelevant differences were found between the distributions of the number of unsampled items for the three non-sampling stations (Figure 5.14). This means, not only the shape of the distribution of the number of consecutive unsampled items relative to the partially skipped station but also the mean value is not affected by the random routes.

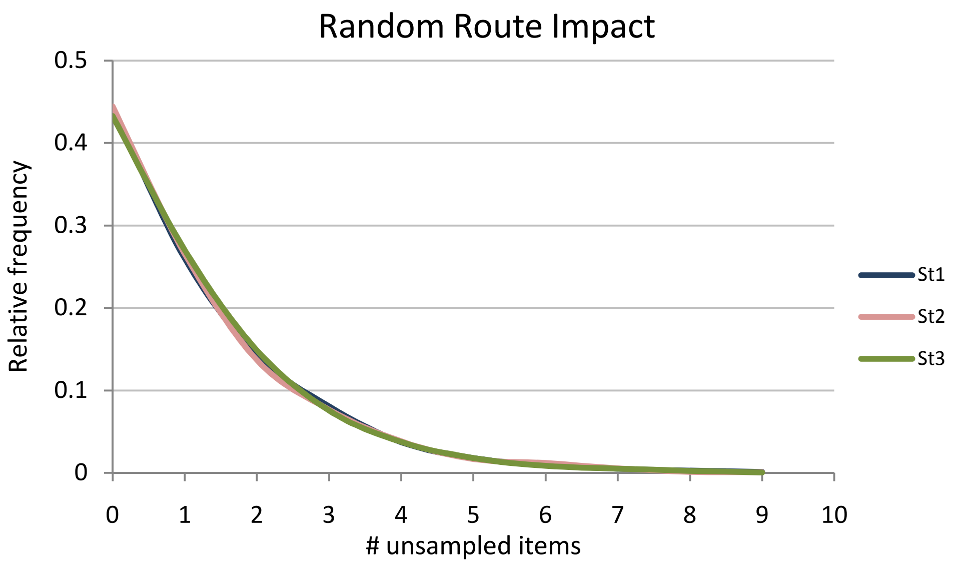


FIGURE 5.14 IMPACT OF A PARTIALLY SKIPPED NON-SAMPLING STATION ON THE DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS IN NON-SAMPLING STATIONS.

On the contrary, the average time between samples in station 2 proves to be twice the time between samples in stations 1 and 3. This suggests that the random routes determine a reduction of the product flow crossing station 2 keeping, on average, the same volume fractions between the two products and the same sampled fraction as observed in the other stations in the segment. This is obviously valid, from a long term perspective, for the randomness of the routing constraints, which are applied in the same fashion independently of the product. For this reason the sampled fraction remains unchanged and, as a consequence, the distribution of the number of unsampled items keeps its mean value unchanged. However, in the skipped station, the inter-arrival times of the two products are obviously increased by the percentage of the flow skipping the station. This causes the increase of the time between samples, which, unlikely the number of unsampled items, is affected by the absolute inter-arrival times.

The shape of the distribution of the number of unsampled items between consecutive samples is still geometric since nothing has changed in terms of flow dynamics. There is still independence in the sequence of successively processed items and the sampled fraction is not affected by the random routes. If the overall sampled fraction had kept changing throughout time, the shape of the distribution would have probably proved to be a combination of geometric distributions characterised by different proportions. This was originally expected; however, once again, the results obtained show the relatively high robustness of the prediction model even in the presence of some relevant variations of the hypothesis on which it was built.

5.4.4 Generalization: basic model

The geometric prediction model mostly bases its validity on both a constant sampling probability and the independence of the events of processing either sampled or unsampled items. Both of these elements should be guaranteed by the inter-arrival dynamics of the different product flows at a machine level; in particular, exponential inter-arrival times should constitute a relevant premise for the independence of the sequence of processed items, due to the memory-less property of the exponential distribution. Moreover, if the arrival/departure process is memory-less, the sampling deterministic pattern, characteristic of the sampling station, is lost; as a result, the sampling probability will be constant for each item. However, there is also a feeling that the disorder level in the system can contribute to this independence effect. The disorder level is intended as both the variability of the time related parameters, which primarily causes the sequence disorder effect, and the randomness relative to the routing patterns, which determines the multiple stream effect. Both these effects contribute to randomise the sequence of items processed consecutively by the machines of the non-sampling stations.

In order to analyse which factors among those cited impact the efficacy of the geometric model in predicting the distribution of the number of unsampled items between consecutive samples, a simulation model of a basic production system was developed in ExtendSim and a few scenarios were simulated. The disorder level was progressively introduced in the scenarios so that the parameters most affecting the distribution could be identified. In particular, the variability of the inter-arrival, processing and waiting times was progressively increased, as shown in Table 5.5.

TABLE 5.5 SCENARIOS CONSIDERED FOR THE ANALYSIS OF THE BEHAVIOUR OF THE BASIC MODEL.

<table>
<tr><th>Scen.</th><th>IAT</th><th>PT1</th><th>QT (imposed)</th><th>PT2</th></tr>
<tr><td>T1</td><td rowspan="4">Det.</td><td rowspan="4">Det.</td><td rowspan="3">No delay</td><td>Det. (<PT1)</td></tr>
<tr><td>T2</td><td>Det. (=PT1)</td></tr>
<tr><td>T3</td><td>Det. (>PT1)</td></tr>
<tr><td>T4</td><td>Det.</td><td>Det. (>PT1)</td></tr>
<tr><td>T5</td><td rowspan="3">Lognormal (CV=0.6)</td><td rowspan="5">Det.</td><td>No delay</td><td rowspan="5">Det. (>PT1)</td></tr>
<tr><td>T6</td><td>Det.</td></tr>
<tr><td>T7</td><td>Det. + Initial level</td></tr>
<tr><td>T8</td><td>Lognormal (CV=1)</td><td rowspan="2">Det. + Inventory increment</td></tr>
<tr><td>T9</td><td>Lognormal (CV=2)</td></tr>
<tr><td>T10</td><td rowspan="2">Lognormal (CV=2)</td><td>Lognormal (CV=0.125)</td><td rowspan="2">Det.</td><td rowspan="3">Det. (>PT1)</td></tr>
<tr><td>T11</td><td rowspan="2">Lognormal (CV=0.625)</td></tr>
<tr><td>T12</td><td>Constant</td><td>No delay</td></tr>
<tr><td>T13</td><td rowspan="4">Lognormal (CV=1)</td><td>Constant</td><td rowspan="4">No delay</td><td rowspan="3">Lognormal (CV=0.176)</td></tr>
<tr><td>T14</td><td>Lognormal (CV=0.187)</td></tr>
<tr><td>T15</td><td>Lognormal (CV=0.312)</td></tr>
<tr><td>T16</td><td>Lognormal (CV=0.5)</td><td>Lognormal (CV=0.529)</td></tr>
<tr><td>T17</td><td>Lognormal (CV=1)</td><td>Lognormal (CV=0.5)</td><td>Lognormal (CV=2)</td><td>Lognormal (CV=0.529)</td></tr>
</table>

The first four scenarios are characterised by deterministic time related input parameters; among other things they allow an analysis of the impact of the differences between the processing time in station 1 (PT1) and station 2 (PT2) on the inventory level. The scenarios from T5 up to T9 introduce inter-arrival time (IAT) randomness, with progressively increasing variability. Different cases for the imposed queuing times are explored at the same time. In particular, the sudden inventory increment introduced in scenarios T8 and T9 is useful to analyse the effect of the inventory level on both the item sequence disorder and the randomisation of routing patterns. The inventory increment is simulated by introducing a given number of items at a given timestamp in the buffer. It is believed that the presence of either an initial or a suddenly increased inventory level can substantially contribute to increase the level of disorder in the system, even though this would be more significant when queuing times are random. Scenarios T10, T11 and T12 focus on PT1 variability; the effect of having deterministic inter-arrival times while the system operates with random processing times is also investigated. Scenarios from T13 to T16 explore the impact of PT2 variability, which is increased along with PT1 variability, on the shape of the distribution of the number of consecutive unsampled items in a non-sampling station. Finally, scenario T17 considers the effect on the same distribution of randomness of all the time related input parameters.

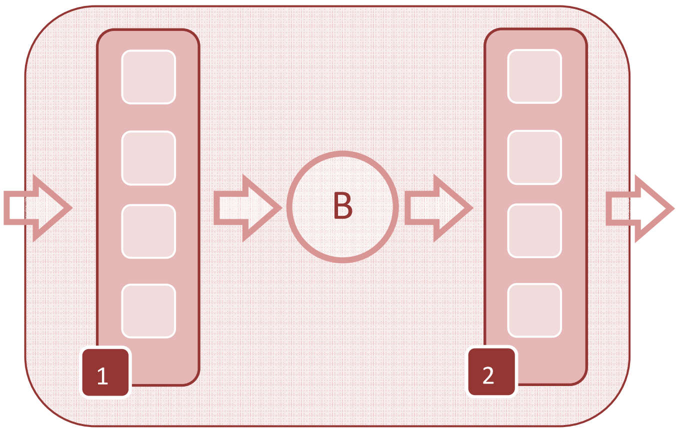


FIGURE 5.15 SIMPLIFIED MODEL STRUCTURE

The simulation model developed consisted of two stations and an intermediate buffer, as shown in Figure 5.15. Both the stations operate with four machines, each of which can only process one item at a time. This constraint applied to the processing capacity could relevantly impact the final results. In fact, even though, the machines in the original model could technically process one item at a time, more than one item could wait within the machine before or after being processed; the waiting times within the machine were considered part of the processing times. As a consequence, the machine could apparently process more than one item at a time and the resulting sequence disorder effect was even higher since the processing times also contributed to it at a machine level. The focus on two stations makes sense since the distance from the sampling station, as previously found, does not relevantly impact the shape of the distribution.

The first scenario analysed, T1, represents a system not affected by time related randomness, at least for the processing times. In fact, deterministic processing times are considered for both the stations and no further delay is imposed in the buffer. Only one product flows serially through the stations. In this case, the sequence of items keeps its order through the line when the ratio between the processing time and the global inter arrival time, which is a kind of utilisation measure, respect some constraints. In the case considered, that is 4 machines in each station, the ratio has to be included between 3 and 4 (scenarios T2 and T3). That guarantees that all the machines in the station are involved in the production. In fact, if that ratio were less than 3, the product flow would be managed by only three machines without any delay caused by an over-utilisation (scenario T4). That obviously happens in the absence of any routing rule, when the decision of the machine where to route an item is left to the software and the strategy of maximising the machine utilisation is implemented. In this condition a biunique relationship therefore exists between Station 1 (St1) and Station 2 (St2) machines and the distribution of the number of the unsampled items is deterministic in both the stations, no matter which is the sampling station.

Randomness was then introduced in the inter-arrival times of items to St1; the processing times were kept constant (scenario T5). The lognormal distribution was used, so that variation of the standard deviation, with no change for the mean value, could be implemented in an easier fashion than in the case of the exponential distribution. With a coefficient of variation of 0.6, only a slight effect of disorder was noticed. This is mainly due to the fact that the inter-arrival times usually prove so high that one of the St1 machines is missed in the ordered sequence of the machines visited. The presence of the buffer along with the fact that the processing time in St2 is greater than in St1 re-establishes a full machine sequence in St2. In this case, the disorder is not intended as the variation in the item sequence since any further time randomness is not imposed to the items after they are introduced in St1. The disorder is more likely due to the multiple-stream effect which causes the loss of the biunique relationships between the machines of St1 and St2. The distribution of the number of consecutive unsampled items is reported in Figure 5.16. A deterministic sampling interval, equal to 3, was alternately set in both the stations, one at a time. Since the model is symmetric, as expected, no relevant change is detectable between the two cases.

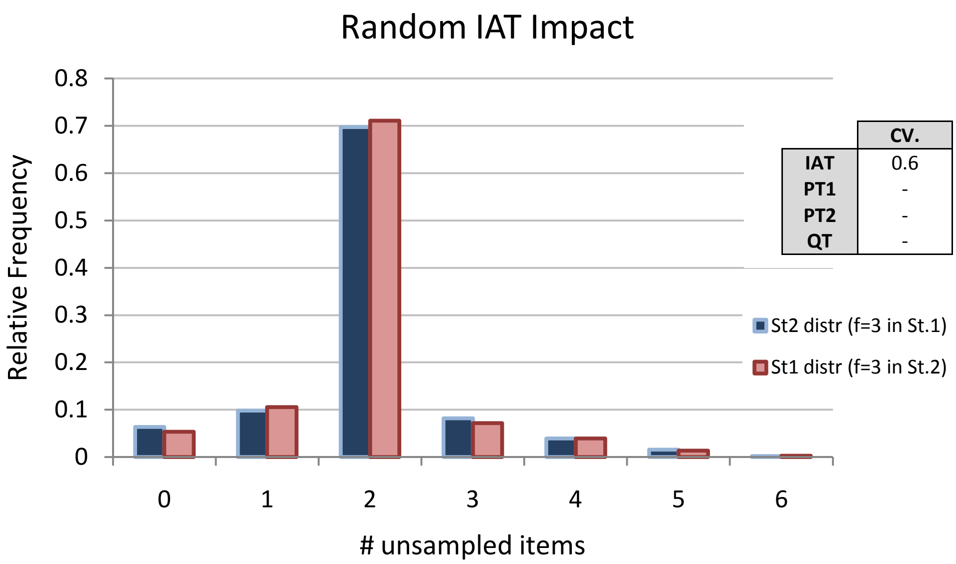


FIGURE 5.16 DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS IN PRESENCE OF RANDOM INTER ARRIVAL TIMES WHEN ITEMS ARE SAMPLED IN ST2 AND ST1 (SCENARIO T5).

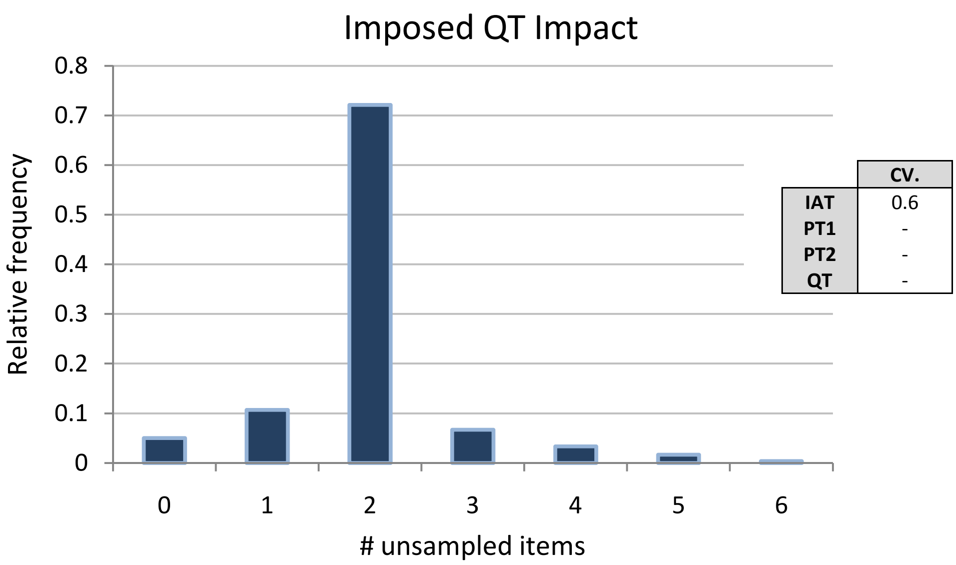


FIGURE 5.17 DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS FOR IMPOSED QUEUING TIMES AND AN INITIAL BUFFER LEVEL GREATER THAN ZERO (SCENARIO T7).

The introduction of a constant imposed delay in the buffer (scenario T6 and T7) has no significant impact on the distribution (Figure 5.17), even in the presence of an initial buffer level. In fact, due to the average machine utilisation, which is close to 85% for St2 machines, the backlog in the buffer can be reduced by temporarily exploiting the remaining machine utilisation; it eventually becomes zero.

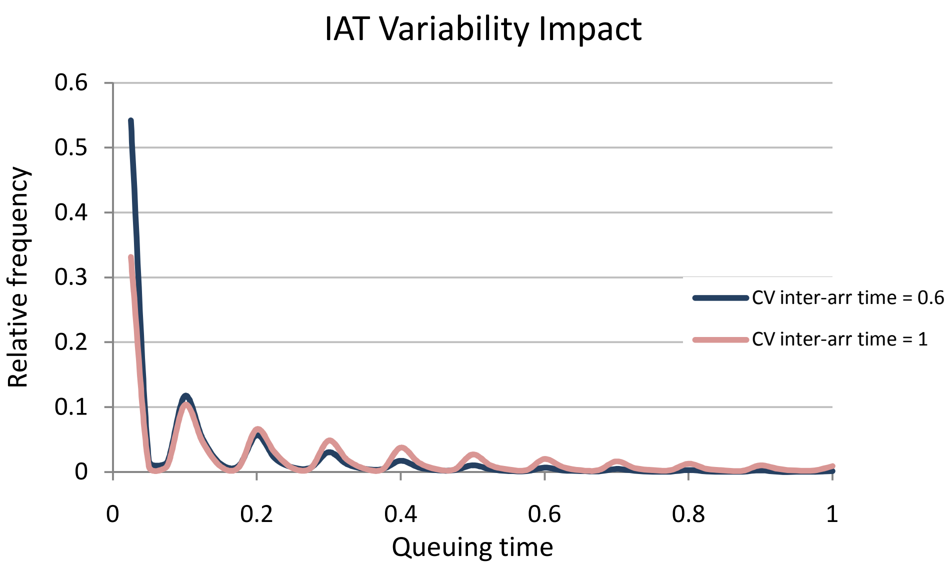


FIGURE 5.18 IMPACT OF THE VARIABILITY OF THE INTER-ARRIVAL TIME ON THE DISTRIBUTION OF THE QUEUING TIME.

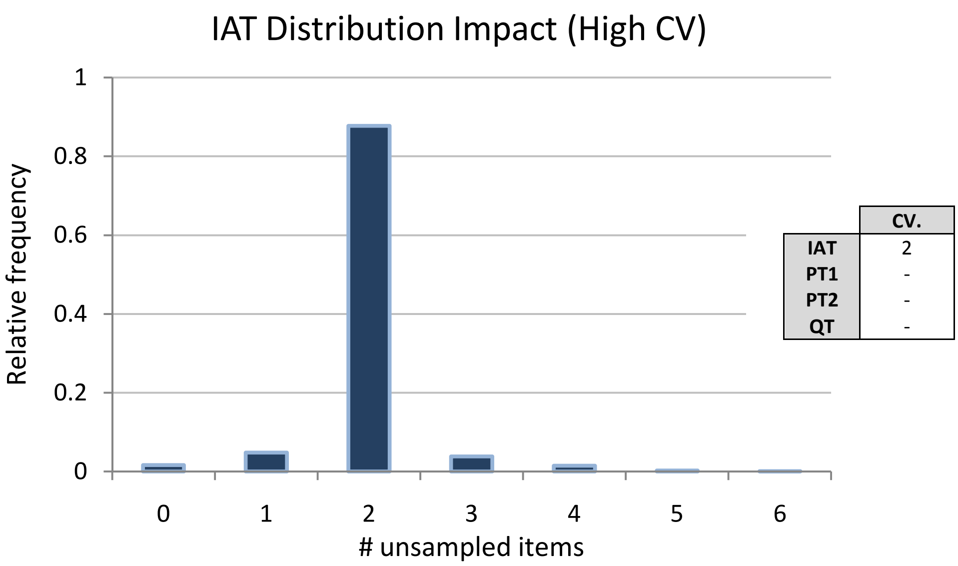


FIGURE 5.19 DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS WHEN THE IAT COEFFICIENT OF VARIATION IS EQUAL TO 2 (SCENARIO T9).

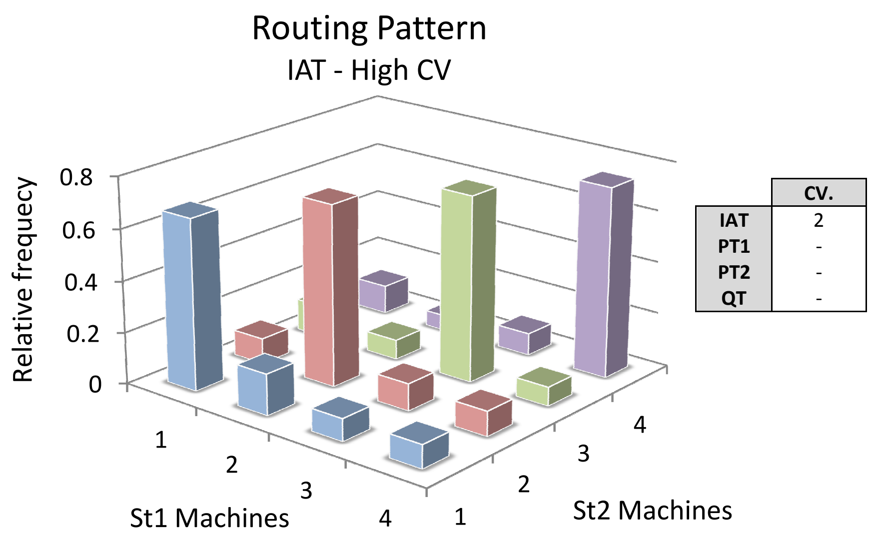


FIGURE 5.20 ROUTING PATTERN BETWEEN ST1 AND ST2 MACHINES FOR SCENARIO T9.

The increase in the coefficient of variation for the distribution of the inter-arrival times (scenario T8) has the only effect of slightly increasing the mean queuing time (Figure 5.18). No effect is registered as regards the distribution of the number of unsampled items, even though, when the coefficient of variation is further increased (CV=2, scenario T9), the distribution seems to become a little peakier on the values of the deterministic sampling interval (Figure 5.19). Preferential routing paths still characterise this scenario at the point that a kind of biunique relationship between St1 and St2 machines is still detectable (Figure 5.20).

When the processing time in St1 is characterised by a lognormal distribution, the effect on the distribution of the number of unsampled items between consecutive samples becomes more significant as the corresponding coefficient of variation of the processing times is increased (scenarios T10 and T11). An increase in the level of the disorder sequence and a randomisation of the routing patterns between St1 and St2 machines (Figure 5.21 and Figure 5.22) are also obtained. The sequence disorder magnitude, measured using Equation 2.2, goes from zero for scenario T9 to 3 and 22 for scenarios T10 and T11, respectively. It is very likely that both these elements constitute the reason why the shape of the distribution of the number of unsampled items progressively gets closer to the geometric distribution. This happens, for example, when the coefficient of variation for the processing times in St1 is equal to 0.625 (Figure 5.23).

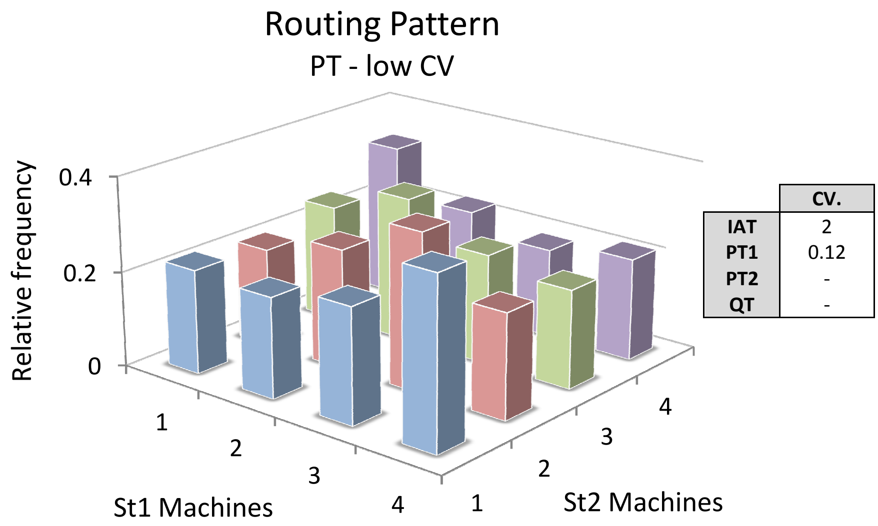


FIGURE 5.21 ROUTING PATTERN WHEN PT1 HAVE CV=0.125 (SCENARIO T10).

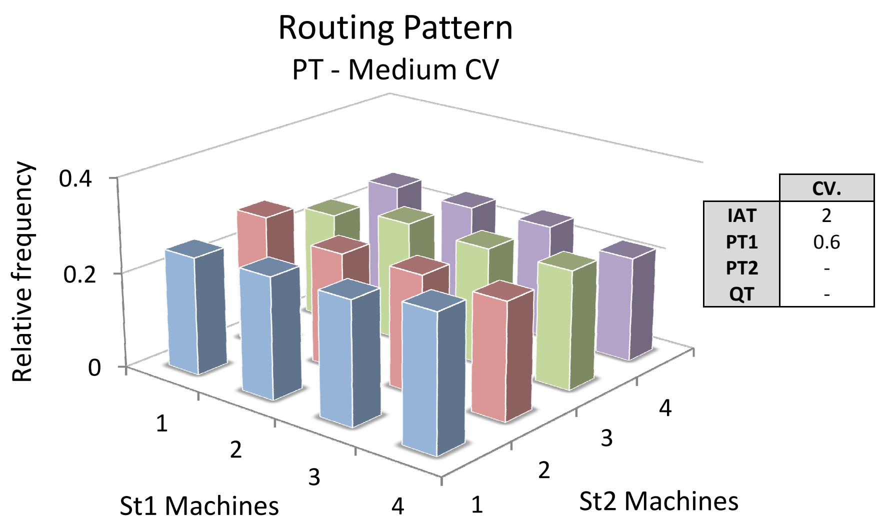


FIGURE 5.22 ROUTING PATTERN WHEN PT1 HAVE CV=0.625 (SCENARIO T11).

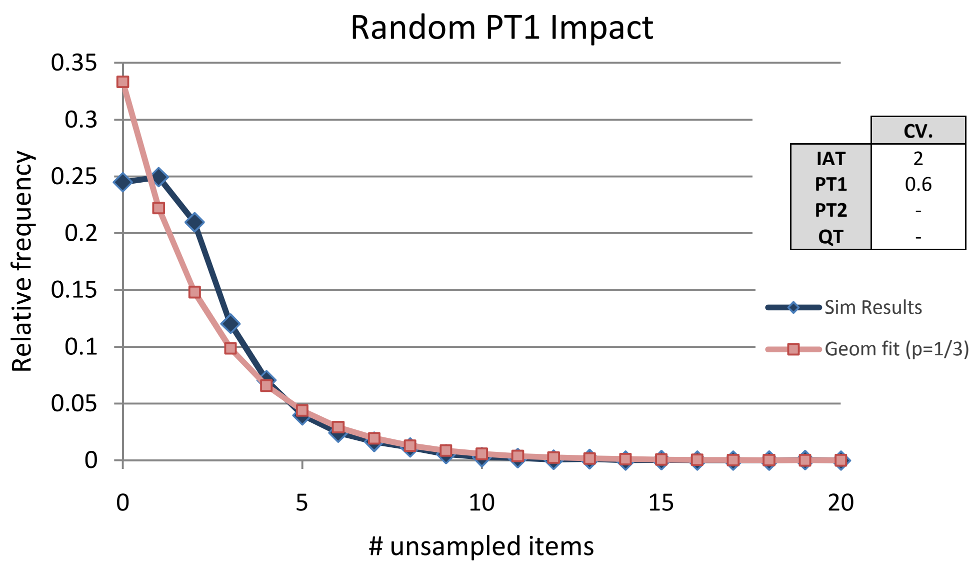


FIGURE 5.23 DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS FOR RANDOM PT1: EFFICACY OF THE GEOMETRIC PREDICTION MODEL (SCENARIO T11).

From the results obtained so far, it is possible to state that, even more than the exponential inter-arrival times, the geometric model needs a reasonably high level of disorder and almost uniform patterns between the machines of the different stations in order to work. This means that the sequence disorder and the multiple stream effects are fundamental factors for the applicability of the model. Whenever the system under investigation is affected by them, the geometric model provides a very accurate prediction for the distribution of the number of unsampled items between consecutive samples.

Keeping the processing time in St1 random with a CV=0.625, the inter-arrival time was set to a constant value (scenario T12), it was found that the impact on the distribution of the number of unsampled items is not relevant. This was easily predictable considering the results previously found for the randomisation of the inter-arrival time in St1. As was affirmed earlier in this chapter, the exponential inter-arrival time presumably constituted the base for the applicability of the geometric model. The results obtained apparently contradict this hypothesis. However, a closer analysis should make clear that, probably, more than the inter-arrival time distribution in station 1, the nature of the problem is conditioned by the inter-arrival distributions in station 2, in this case, and in the stations comprised between the first/last one in the segment and the sampling station, for more general cases. For these stations, the sequence disorder and the multiple stream effects make the inter-arrival time distribution tend to an exponential distribution (Figure 5.24). Analogously, the nature of the departure process from a station could play a major role for the validity of the geometric model.

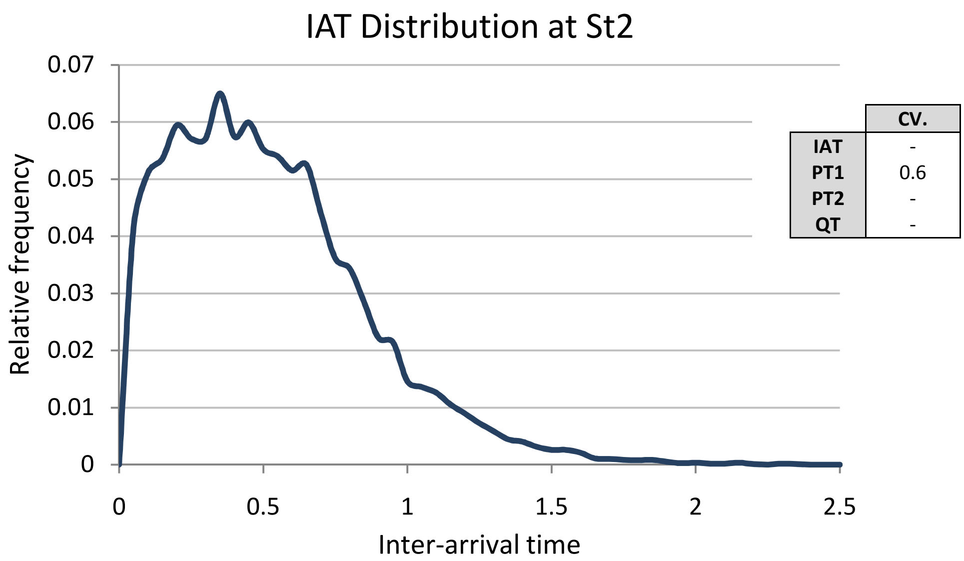


FIGURE 5.24 INTER-ARRIVAL TIME DISTRIBUTION AT ST2.

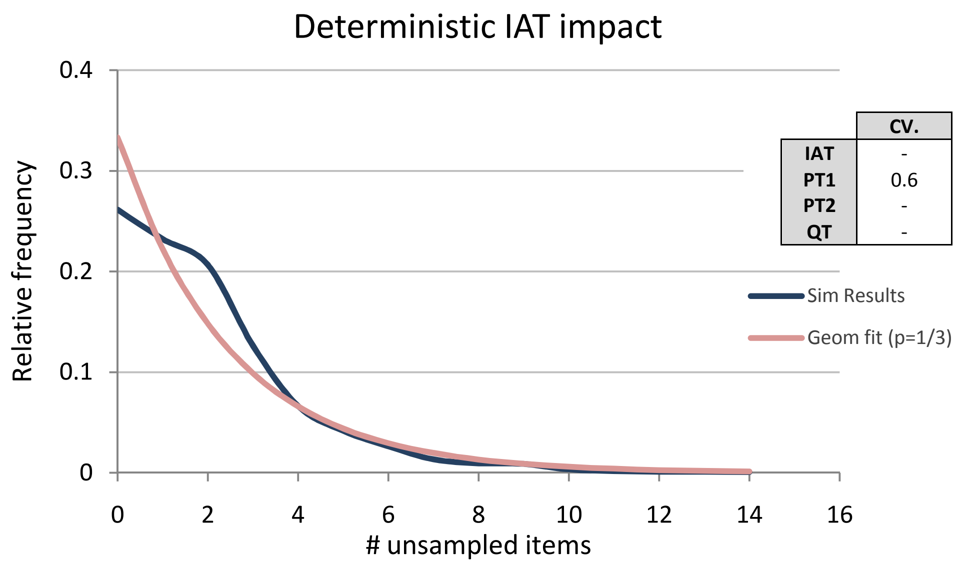


FIGURE 5.25 DISTRIBUTIONS OF THE NUMBER OF UNSAMPLED ITEMS FOR SCENARIO T12: COMPARISON WITH THE GEOMETRIC PREDICTION MODEL.

The curves in Figure 5.25 confirm the goodness of fit of the geometric distribution for the upper tail of the distribution of the number of unsampled items; the greatest error is registered for the region of theoretical sampling interval to the detriment of the probability frequency at 0 unsampled items.

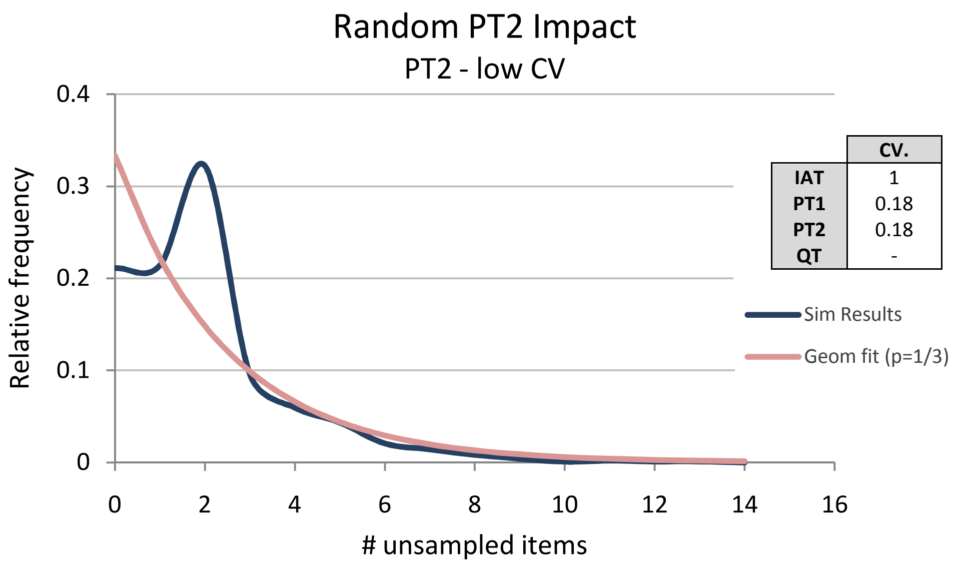


FIGURE 5.26 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS FOR RANDOM PROCESSING TIMES (SCENARIO T14): COMPARISON WITH THE PREDICTION MODEL.

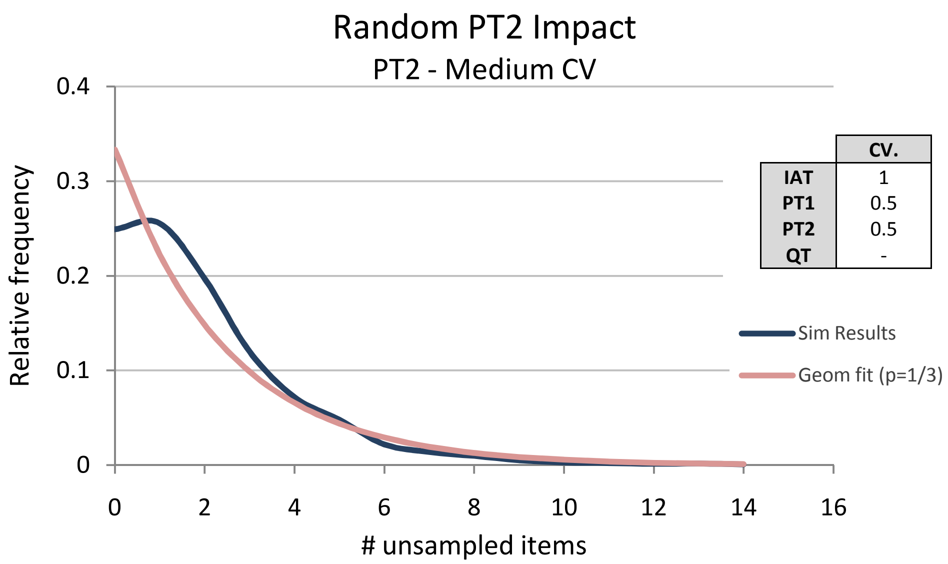


FIGURE 5.27 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS FOR RANDOM PROCESSING TIMES (SCENARIO T16): COMPARISON WITH THE PREDICTION MODEL.

An analogous compensation effect between the probability frequencies corresponding with 0 and 2 (3 being the sampling interval) unsampled items is obtained when the processing times of both the stations are log-normally distributed with a coefficient of variation approximately equal to 0.18 (scenario T14) (Figure 5.26). In this case, random inter-arrival times have been considered (CV=1). When the coefficient of variation of the processing times for both the stations is increased (CV=0.5) (scenario T16) the distribution becomes smoother (Figure 5.27) and much closer to the geometric distribution. In order to avoid repetitiveness, results coming from scenarios T13 and T15 are not shown since similar to the results of the other scenarios illustrated above.

Finally, random imposed queuing times were introduced in the system so that all the time parameters would be random variates (scenario T17). The resulting distribution of the number of unsampled items is somewhat smoother, which means it is characterised by less dispersion (Figure 5.28). The geometric distribution seems to fit better than in the previous cases. However, a discrepancy between the frequency values for 0 and 2 unsampled items is still detectable.

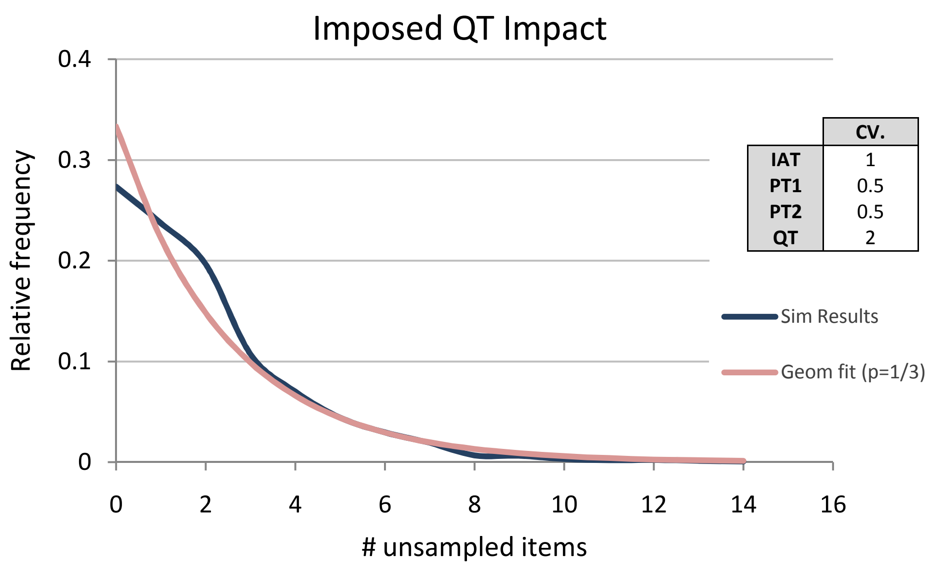


FIGURE 5.28 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS AFTER THE INTRODUCTION OF IMPOSED QUEUING TIMES: COMPARISON WITH THE PREDICTED DISTRIBUTION.

The suspected attitude of the experimental distribution to depart from the geometric distribution in the nearness of the deterministic sampling interval was investigated by changing the sampling parameter and keeping everything else the same. The graphs in

Figure 5.29 and Figure 5.30 confirm this suspicion. Mainly in Figure 5.29, where a sampling interval of 5 was adopted, the effect of concentrating the probability frequency in the nearness of 4 to the detriment of smaller values is clear.

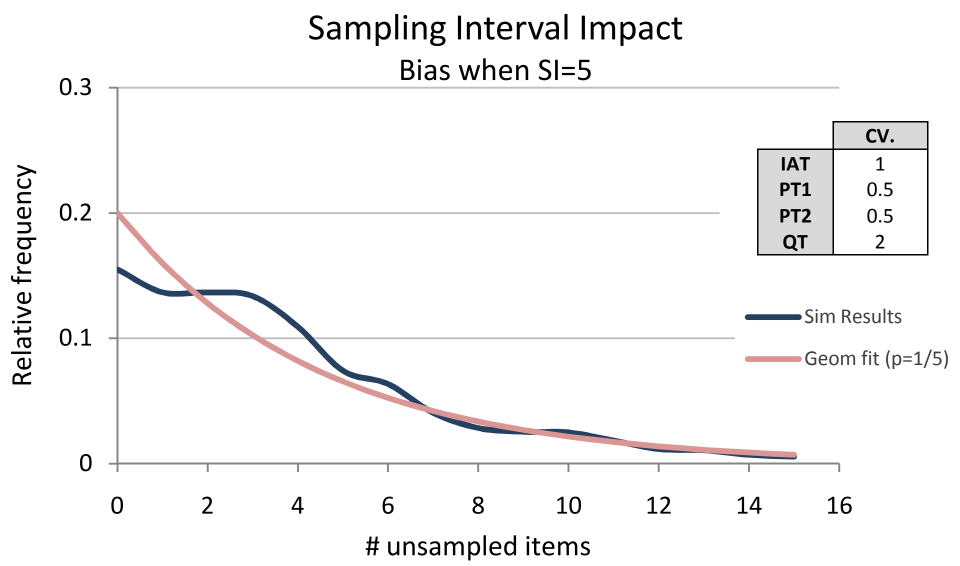


FIGURE 5.29 DISTRIBUTIONS OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: BIAS IN THE NEARNESS OF THE SAMPLING INTERVAL (SI=5).

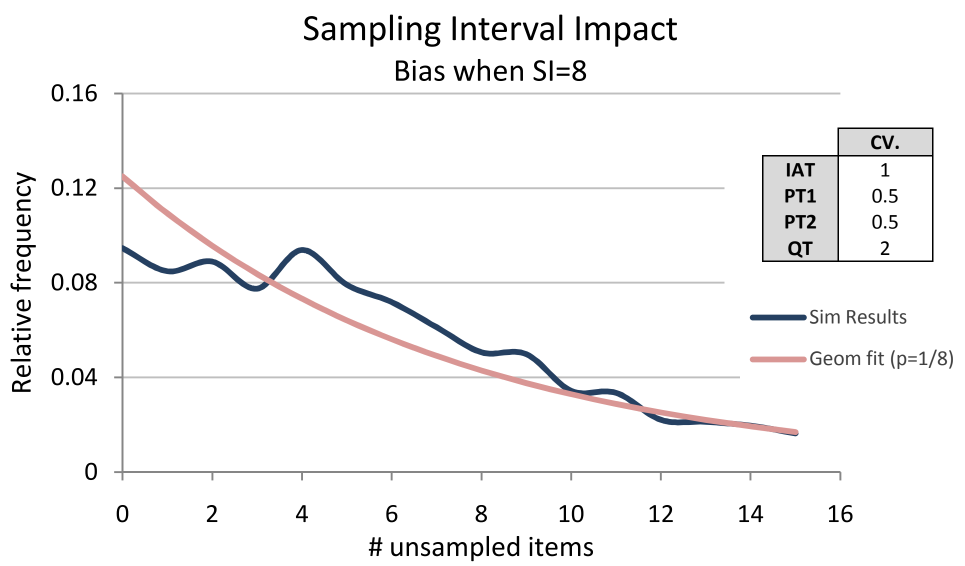


FIGURE 5.30 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: BIAS IN THE NEARNESS OF THE SAMPLING INTERVAL (SI=8).

Evidently, when the sampled fraction becomes smaller, as in the case of Figure 5.30 where a sampling interval equal to 8 was considered, this effect is spread across a larger domain. This presumably cause a better fit when very small sampled fractions are considered, as happens for the case analysed in the original simulation model.

Another observation can be inferred by the graphs in Figure 5.29 and Figure 5.30. If a system is characterised by a low level of sequence disorder, as a system with a simple structure usually is, the sampling interval of the monitored product will impact the distribution of the number of consecutive unsampled items in a non-sampling station machine in a more significant way than in a system characterised by a high level of disorder. Since the deviation of the actual distribution of the number of consecutive unsampled items from the geometric model occurs in the nearness of the sampling interval and interests a region which gets wider as the sampling interval increases, opting for smaller sampling intervals also means that the distribution becomes more right-skewed. This causes immediate benefits in terms of quality risk since the probability of observing a high number of unsampled items reduces.

Finally, it has to be noted that the decision to model the machine so that only one item at a time could be processed certainly causes a relevant reduction of the sequence disorder effect. In fact, items are forced to keep the same sequence in and out of a machine. The sequence disorder for the particular scenarios investigated in this section is mainly due to the randomness of the queuing times. This could be another reason of the small distortion of the actual distribution of the number of unsampled items in the nearness of the deterministic sampling interval relatively to the case with no processing overlapping in the operating machines.

5.4.5 Stochastic approach

The analytical approach used so far has involved a pure statistical analysis of the simulation results, the individuation of algebraic functions for the data description and some basic knowledge of probability theory. That worked quite well and provided prediction models which can be easily applied. However, considering the nature of the problem, another approach can be followed. In fact, when different products flow through a parallel production system with random inter-arrival and processing times at each station, the type of the item moving out from a particular tool in any station, at any time *t*, can be considered a discrete random variable whose number of states and their corresponding probabilities are determined by the product mix. The family of these random variables, as time goes by, can be studied as a stochastic process. With good approximation, this can be considered a Markov process since the occurrence of each state depends at most only on the immediately previous state and not on the sequence of several preceding states.

The number of states of the random variable, which are at least equal to the number of products which cross a station, can be increased if a further distinction between sampled and unsampled items is made. Considering the sampling strategy analysed in this thesis, the sequence of sampled and unsampled items for any monitored product is deterministic only for the sampling station. In fact, the sequence disorder and the multiple stream effects turn that ordered sequence into a memory-less, random sequence in any machine of the stations upstream/downstream from the sampling station from both a single product and a global product flow perspective. In this case, splitting the state associated with one particular product into the sampled and the non-sampled state simply means partitioning the probability of the original state into two complementary probabilities based on the sampling interval of that product.

Modelling this system as a Markov chain should easily allow the evaluation of the steady-state probabilities which are the probabilities associated with any state of the system in the long term. Their inverse represents the mean return time to the same state which is particularly interesting from a quality risk point of view. In fact, the sum of the steady-state probabilities relative to the sampled states of all the monitored products represents the probability that a processed item could be measured at the end of the line. The inverse of this probability estimates the average number of items between consecutive samples in any machine of a station, which is the performance measure also predicted by Equation 5.5. The comparison of the results obtained using the stochastic approach with the predictions coming from the deterministic prediction formula could prove interesting.

In order to see how stochastic theory can help in the evaluation of the quality risk, let us consider an example. A station is crossed by three products arriving at the station with exponentially distributed times. The mean inter-arrival times are 2, 4 and 3 time units for product A, B and C, respectively, which gives an approximated volume fraction of 0.46, 0.23 and 0.31 for the three products. Product A and B are monitored with a sampling interval of 3 and 4, respectively. This means that one third of product A items are sampled and the remaining two thirds are unsampled. Hence, the 46% product A volume fraction can be further divided into 15% and 31% which represent the sampled fraction and the unsampled fraction of product A, respectively. The same approach can be followed for product B; it makes no sense doing the same for product C, since it is not monitored.

The system can be then described in terms of a Markov chain with a finite number of states, 5, in this case, and a transition matrix which can be built based on the volume fractions. Indicating with subscript *u* the unsampled fraction and with subscript *s* the sampled fraction, the transition matrix for the case described is the following:

$$
\begin{array}{cc|ccccc}
 & & & & \text{Destination} & & \\
 & & A_s & A_u & B_s & B_u & C \\
 & A_s & 0 & 0.46 & 0.06 & 0.17 & 0.31 \\
 & A_u & 0.15 & 0.31 & 0.06 & 0.17 & 0.31 \\
\text{Source} & B_s & 0.15 & 0.31 & 0 & 0.23 & 0.31 \\
 & B_u & 0.15 & 0.31 & 0.06 & 0.17 & 0.31 \\
 & C & 0.15 & 0.31 & 0.06 & 0.17 & 0.31
\end{array}
\tag{5.9}
$$

The zeros in matrix (5.9) show that when the system is in a sampled item state it can not immediately return to itself. This way of modelling particularly fits the behaviour of the system in a sampling station, since the sequence of sampled and unsampled items is clearly defined by the sampling interval. Suspending for the moment the analysis of matrix (5.9), in a non-sampling station, this constraint can be loosened. In fact, the sequence disorder effect could change the sequence of processed items so that two sampled items of the same product can consequently move out of a machine. In this case, the transition matrix will appear as follows

$$\begin{array}{cc|ccccc} & & & & \text{Destination} & & \\ & & A_s & A_u & B_s & B_u & C \\ \hline \text{Source} & A_s & 0.15 & 0.31 & 0.06 & 0.17 & 0.31 \\ & A_u & 0.15 & 0.31 & 0.06 & 0.17 & 0.31 \\ & B_s & 0.15 & 0.31 & 0.06 & 0.17 & 0.31 \\ & B_u & 0.15 & 0.31 & 0.06 & 0.17 & 0.31 \\ & C & 0.15 & 0.31 & 0.06 & 0.17 & 0.31 \end{array} \tag{5.10}$$

All the rows in the matrix are the same, which means that the probability of reaching a particular state does not actually depend on the initial state of the system. The evaluated probabilities are, then, the steady-state probabilities based on which the average number of unsampled items between consecutive samples can be calculated. This result agrees with what was previously found with the prediction model described by Equation 5.5.

In fact, using the steady-state probabilities from the transition matrix, it results that the system is under a sampled state for a fraction of:

$$A_s + B_s = 0.15 + 0.06 = 0.21 \tag{5.11}$$

which means that to return to a sampled state it is necessary to wait, on average, for a number of items given by the following formula:

$$mean\,recurrence\,time = \frac{1}{A_s + B_s} = \frac{1}{0.21} = 4.76 \tag{5.12}$$

Since this result include the sample, the average number of unsampled items between consecutive samples is then equal to 3.76, that is 4.76-1, which is the same value obtained by the formula in Equation 5.5. The results obtained show that the prediction model developed on a deterministic base has complete support from a stochastic theory point of view as well.

For completeness sake, the same procedure followed for matrix (5.10) was also applied to matrix (5.9). The impact of the use of matrix (5.9) on the goodness of fit of the geometric model for the prediction of the distribution of the number of unsampled items between consecutive samples was also investigated. It is worth remembering that this approach is more suitable for describing the behaviour of the system in a sampling station.

The system described by a transition matrix of the likes of matrix (5.9) is actually an ergodic Markov chain, which means all its states are recurrent and aperiodic. A state is recurrent when it has opportunities to be revisited from other states; it is aperiodic when it is not periodic, which means a return to the state is not constrained to any period of time.

For an ergodic chain, the steady-state probabilities π_i can be calculated from the equations

$$\begin{aligned} &\pi = P\pi \\ &\sum_i \pi_i = 1 \end{aligned} \tag{5.13}$$

where P is the transition matrix. The equations in the set $\pi = P\pi$ are not linearly. One of the equations is redundant and substituted by the normalising equation, $\sum_i \pi_i = 1$. Solving this set of equations gives the same results as multiplying the transition matrix by itself for a number of times until no change is detectable in the result of the multiplication. Once the steady-state probabilities are obtained, the same procedure used before can be applied for the calculation of the average number of unsampled items between consecutive samples.

For the example described before, the set of equations in Equation 5.13 gives the following steady-state probabilities

$$\begin{array}{ll} A_s & 0.13 \\ A_u & 0.33 \\ B_s & 0.05 \\ B_u & 0.18 \\ C & 0.31 \end{array} \tag{5.14}$$

which means that the system is in a sampled condition with an 18% probability (A_s + B_s). The number of unsampled items between samples is now increased to 4.55 ($\frac{1}{A_S+B_S}-1$).

The fact that the probability relative to the sampled states is reduced was quite expectable. In fact, the sampled states can not return to themselves; they can be reached only from other states. The sum of the probabilities relative to a product, as before, is exactly equal to the volume fraction of that product; however, in this case, there is a kind of imbalance between the states associated with a product and the partitioning of the steady-state probabilities does not perfectly follow the sampling intervals.

There is only a trivial case where the two matrix building approaches converge to the same results. It is the case when the sampling intervals are 2 for both the products. Considering the case where only products A and B flow through the station with the same inter-arrival times as before, the transition matrix with sampling intervals equal to 2 is given by

$$\begin{array}{cc|cccc} & & \multicolumn{4}{c}{\text{Destination}} \\ & & A_s & A_u & B_s & B_u \\ \hline \multirow{4}{*}{\text{Source}} & A_s & 0 & \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \\ & A_u & \frac{2}{3} & 0 & \frac{1}{6} & \frac{1}{6} \\ & B_s & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ & B_u & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{array} \tag{5.15}$$

Since a sampling interval of 2 means that alternate items are measured, the sampled and unsampled conditions of a product alternate. The matrix reacquires a perfect balancing between the states, since it happens that the unsampled states can not immediately return to themselves. In fact, the application of equations 5.13 produces the following results

$$\begin{array}{cc} A_s & \frac{1}{3} \\ A_u & \frac{1}{3} \\ B_s & \frac{1}{6} \\ B_u & \frac{1}{6} \end{array} \tag{5.16}$$

which are the exact same as the results immediately obtained building the matrix by means of the approach relative to a non-sampling station

$$\begin{array}{ccccc} & A_s & A_u & B_s & B_u \\ A_s & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ A_u & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ B_s & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ B_u & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \end{array} \tag{5.17}$$

It is quite intuitive that if any other unmonitored product was present, nothing would change as long as all the monitored products have a sampling interval equal to 2.

Excluding this last trivial case, different sampling intervals were considered for the scenario relative to matrix (5.15). The steady-state probabilities found were then used to evaluate the parameter of the geometric model, so that the distribution of the number of unsampled items in a non-sampling station can be predicted. The distributions predicted using the sampled fractions obtained from the stochastic analysis were compared with simulation results; the comparisons show that the prediction is relatively good with cumulative absolute error ranges always less than 10%. However, the performances of this approach are always poorer than the ones obtained using Equation 5.5. This is particularly evident for the high sampling frequencies, most likely because the sampled fraction is large and comparable with the unsampled fraction. When the magnitude of the sampled fraction reduces, the impact of the steady-state probability approach on the original probabilities is less evident and the error consistently reduces and gets closer to the error associated with the algebraic approach, that is the approach based on the AP formulae.

The results reported in Figure 5.31 refer to a scenario characterised by the presence of two monitored products, A and B. Product A has an inter-arrival time equal to 2 [time unit/item] and its sampling interval is set to 2; product B has inter-arrival time equal to 4 [time unit/item] and a sampling interval set to 3. The comparison proposed in Figure 5.31 clearly shows that the algebraic approach performs much better than the Markov chain approach. The cumulative error of the latter reaches 10.5%, whereas, the error of the former is 3.43%. As a consequence of the reduction in the steady-state probabilities of the sampled states, the Markov chain approach predictions always underestimate the lower tail of the distribution of the number of unsampled items in contrast to what happens with the algebraic approach.

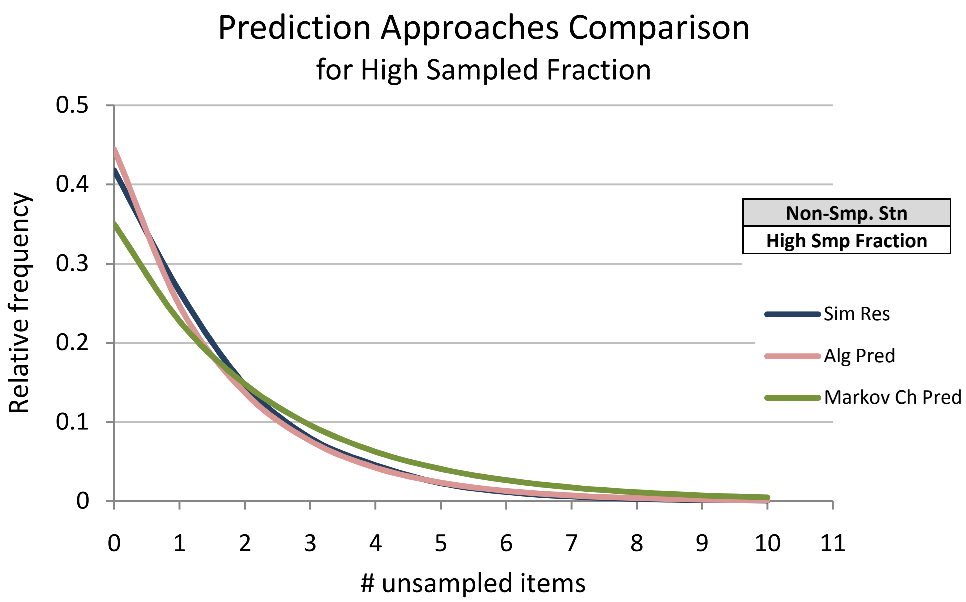


FIGURE 5.31 COMPARISON BETWEEN THE ALGEBRAIC AND THE STOCHASTIC APPROACHES FOR THE PREDICTION OF THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS.

As discussed earlier, for larger sampling intervals the performances of the Markov chain predictions clearly improve, as shown in Figure 5.32, where both products are sampled with an interval equal to 4. The algebraic approach still outperforms the

Markov chain approach in terms of cumulative absolute error ranges, which are 4.7% and 5.5%, respectively. However, the Markov chain approach has better performance in terms of absolute error ranges, which are 3.61% and 2.18%, respectively.

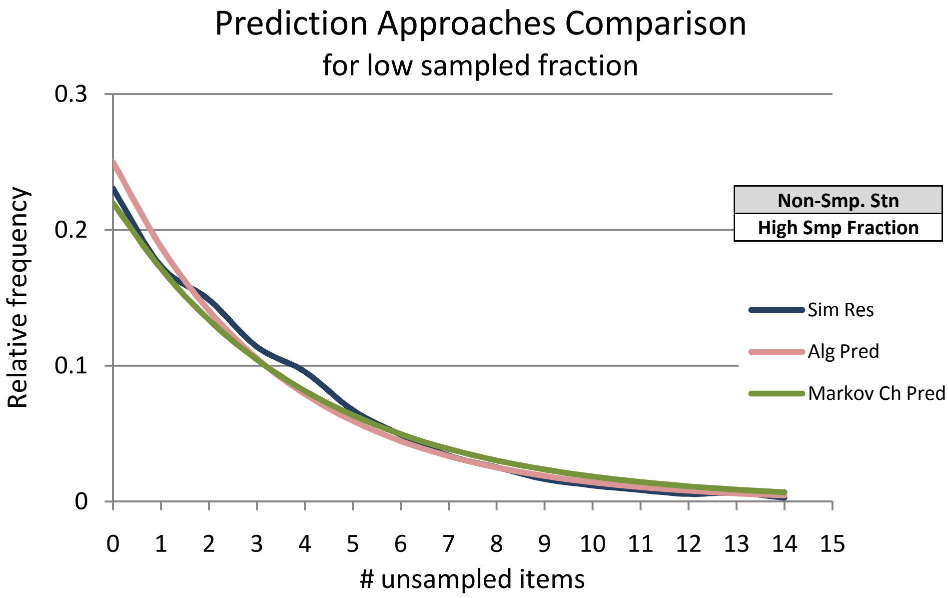


FIGURE 5.32 COMPARISON BETWEEN THE ALGEBRAIC AND THE STOCHASTIC APPROACHES FOR LOW SAMPLED FRACTIONS.

The tendency of the Markov chain approach to underestimate the distribution of the number of unsampled items makes the minimum cumulative error quite large in comparison with that obtained from the deterministic approach for which the associated geometric distribution has quite well balanced underestimated and overestimated areas with respect to the simulation results.

Finally, a similar smoothing effect as obtained with the larger sampling intervals is determined by the introduction of unmonitored products. The unmonitored items reduce the monitored volume fraction, thus reducing the sampled fraction as well. In absolute terms, the Markov chain procedure has very little impact on the smaller fractions, that is, the steady state probabilities calculated are smaller than the volume fractions in input. As a consequence, the Markov chain approach and the algebraic approach produce very similar distributions. An example of that is provided in Figure 5.33. Here the unmonitored product represents the 91.1% of the volume fraction and the effect of the Markov chain approach on the sampled steady-state probability produces some benefits on the prediction of the distribution of the number of unsampled items between consecutive samples. In fact, the cumulative absolute error ranges for the Markov chain approach, 4.31%, is a bit less than the error associated with the mathematical approach, 4.55%. The better fit of the geometric prediction based on the stochastic approach is also confirmed by the Person's chi square test. The P-value for the geometric prediction based on the stochastic approach ($2.99*10^{-2}$) is slightly greater than the P-value for the geometric prediction based on the deterministic approach ($2.26*10^{-2}$).

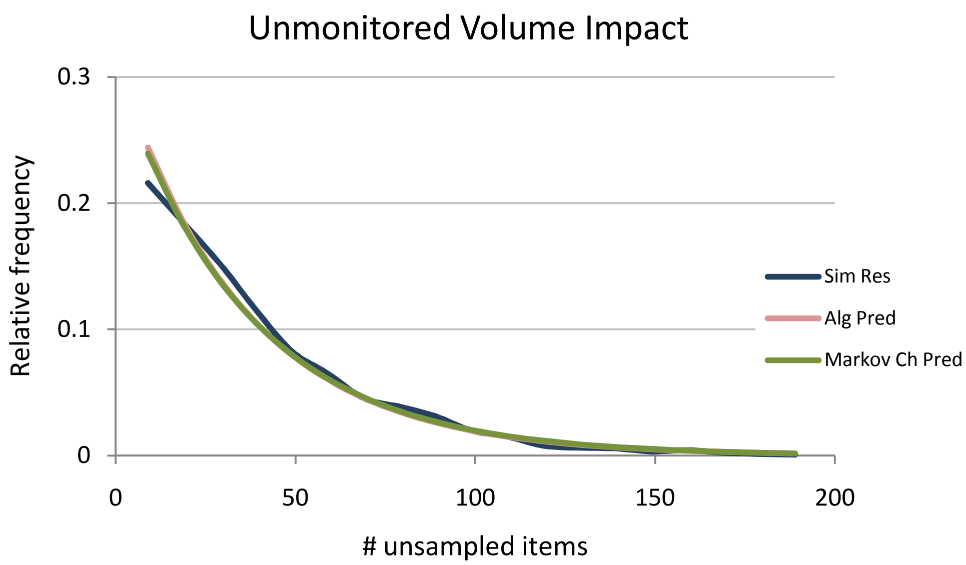


FIGURE 5.33 COMPARISON BETWEEN THE ALGEBRAIC AND STOCHASTIC APPROACH FOR THE PREDICTION OF THE DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS.

In conclusion, the main finding in the analysis of the average number of unsampled items between consecutive samples from a stochastic perspective is that the application of a Markov chain approach to the problem of finding the average number of items between samples is equivalent to the algebraic approach previously shown. Moreover, there exists a way to build the transition matrix, which would more suitably model the behaviour of the system in a sampling station. This usually provides slightly worse results than the algebraic approach, unless very small sampled fractions are considered. In this case, the sampled fraction calculated based on the sampled steady-state probabilities proves a bit smaller than the actual sampled fraction. As a consequence, the parameter of the geometric prediction model is reduced. This is beneficial to the prediction of the distribution of the number of unsampled items whose lower tail is usually overestimated by the algebraic approach.

5.5 Sampling station case

When a sampling station is considered, the sequence disorder and the multiple-stream effects are no longer relevant for the analysis of the distribution of the number of unsampled items. In fact, being the station where the sampling decision is made on a deterministic basis at the moment when items move out of a machine, the randomness of both the time related parameters and the routing patterns do not immediately affect the dynamics behind the definition of the distribution of the number of consecutive unsampled items. Indeed, the sampling station can be considered the source of the quality information, no matter where it is located in the production segment which it is meant to monitor. This might lead one to consider that the sampling station case is easier to investigate than the non-sampling station case.

However, only one trivial case exists. That is the scenario with only one product crossing the station; if the product is monitored with sampling interval f, the distribution of the number of unsampled items between consecutive samples banally degenerates to a one-value distribution with probability equal to 1. The only value is obviously equal to f-1.

Excluding this case, complexity factors, partially different from those impacting the non-sampling station scenarios, intervene in the development of the prediction model for the distribution of the number of consecutive unsampled items. These factors, which are mainly traceable to the product flow complexity, will be gradually introduced in the model and their impact on the distribution of the number of unsampled items will be progressively explored. With this aim the scenarios characterised by one monitored product (Product 1 (Pr.1)) and unmonitored product flow will be first investigated. Then the introduction of a second monitored product (Product 2 (Pr.2)) will be considered, with and without the presence of an unmonitored product flow. As noted in the previous sections, unmonitored products can either cross the single station or flow serially through the segment; the nature of the unmonitored flow does not impact the characteristics of the resulting distribution of the number of unsampled items between consecutive samples.

5.5.1 One monitored product + unmonitored flow case

The first complexity factor introduced in the analysis of the distribution of the number of unsampled items between consecutive samples is the presence of an unmonitored product flow into the sampling station. This turns the number of consecutive unsampled items into a random variable. In fact, due to the randomness of the inter-arrival times, there is no theoretical limit to the number of unmonitored items which can be produced between two consecutive monitored items. Indeed, for the same reasons highlighted for the non-sampling station case, the succession of the items moving out from a machine can be considered a geometric process. This means that the sequence of monitored and unmonitored items registered immediately downstream of a machine is definitely random and based on the proportions of the volume mix.

A deterministic element is still associable with this particular scenario. In fact, the monitored flow is made up of one product; hence, from a monitored flow perspective, the sampling interval coincides with the sampling interval of the only monitored product flowing through the segment. So, while processing an unmonitored item does not trigger any particular event, it is necessary to keep the count of the number of monitored items processed by a machine, since the r^{th} item will be chosen as a sample, when r is the sampling interval of the monitored product type.

Once the problem of finding the distribution of the number of unsampled items is formulated in this fashion, its solution is easy to find among the most common known discrete distributions. In fact, the definition of the negative binomial distribution, if opportunely interpreted, seems to match the problem thesis. A negative binomial distribution describes the number of failures before the r^{th} success in a sequence of independent Bernoulli trials with probability p of success (See Appendix D). From the perspective of the investigated problem, a failure is an unmonitored item, as a consequence, a success would be a monitored item and, in particular, the r^{th} success is the chosen sample; the independent Bernoulli trials are obviously the items processed, whose sequence is actually independent, and, finally, the probability of success represents the monitored volume fraction. As requested from the negative binomial distribution hypotheses, the probability of success is constant for each event observed. In other words, processing an item either monitored or unmonitored is the realisation of a Bernoulli event; as a consequence the distribution of number of unmonitored items processed before observing a monitored item should follow a geometric distribution. Since before choosing a sample *r* monitored items have to be processed, *r* geometric distributions have to be summed and a negative binomial distribution is obtained. Following this logic, at least theoretically, the number of unsampled items between consecutive samples in a sampling station crossed by one monitored product type and an unmonitored flow should follow a negative binomial distribution, with parameters given by the sampling interval of the monitored product and the monitored volume fraction, shifted by the sampling interval minus one.

The need for shifting the negative binomial distribution is a direct consequence of the definition used for it. In fact, it exclusively takes into account the number of failures, which are the unmonitored items, before the r^{th} success. However, in the number of unsampled items, the monitored items between two consecutive samples should also be included. These items can be formally considered successes, since they contribute to triggering the sampling choice; however, from a quality viewpoint they are still items at risk since they can not be used to spot quality failures and, hence, can only be passive carriers of it.

It is worth noting that some authors, as for example Montgomery [34], define the negative binomial distribution as the distribution which describes the number of independent trials, no matter whether successful or not, before the r^{th} success; according to this definition, the domain of the distribution presents as its lower limit the value r-1. Obviously, in this case, no shift would be needed.

In order to assess the efficacy of the negative binomial distribution in modelling the number of unsampled items between consecutive samples for the scenario investigated in this section, the simulation experiments conducted for the non-sampling station case and reported in (p. 142) proved to be useful for investigating the robustness of the prediction model to variations of its parameters.

Figure 5.34 and Figure 5.35 show the results obtained for two scenarios characterised by very high monitored volume fractions. The accuracy of the prediction model is very high so that the representation of the predicted distribution and the actual distribution as curves would hardly reveal eventual prediction errors. Then, the representation of the distributions by means of histograms was preferred.

The application of the Pearson's chi square test confirmed the high goodness-of-fit of the negative binomial model; the P-values obtained for the different scenarios ranged from 0.03 (for a case with one degree of freedom) to 0.91 (for 16 degrees of freedom). The choice of the number of classes was based on the number of available samples and on the simulation results obtained; until the absolute frequency of samples in a class was consistently over 5 items, additional classes were considered. The width of each class was chosen equal to one, unless merging consecutive classes was necessary to increase the number of observations in a class (more than 5). Being the number of parameters of the negative binomial distribution equal to 2, the correction to the degrees of freedom for the chi square test was equal to 3.

The maximum absolute error is less than 0.55% for the scenario in Figure 5.34 and smaller than 0.40% for the scenario in Figure 5.35. These scenarios present an average absolute error of 0.11% and 0.14% respectively; the average absolute error was calculated based on the first 15 points of the distribution. The higher maximum absolute error registered for the first scenario with respect to the second scenario is a consequence of the more limited range and, therefore, the higher relative frequencies which interest the first distribution. In relative terms, that wouldn't happen. It is opportune to note that the absolute error is a reasonable choice as an error measure, since the distributions are normalised and the values involved are usually very small, in particular for the upper tail of the distributions. That would result in very high average relative errors which would mislead the goodness-of-fit analysis.

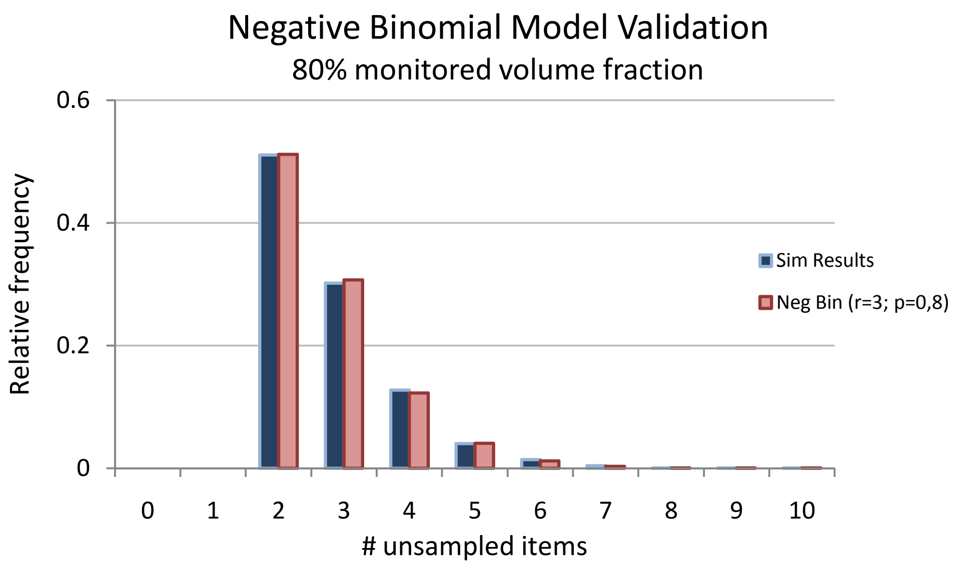


FIGURE 5.34 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: VALIDATION OF THE NEGATIVE BINOMIAL PREDICTION MODEL FOR R=3 AND P=0.8.

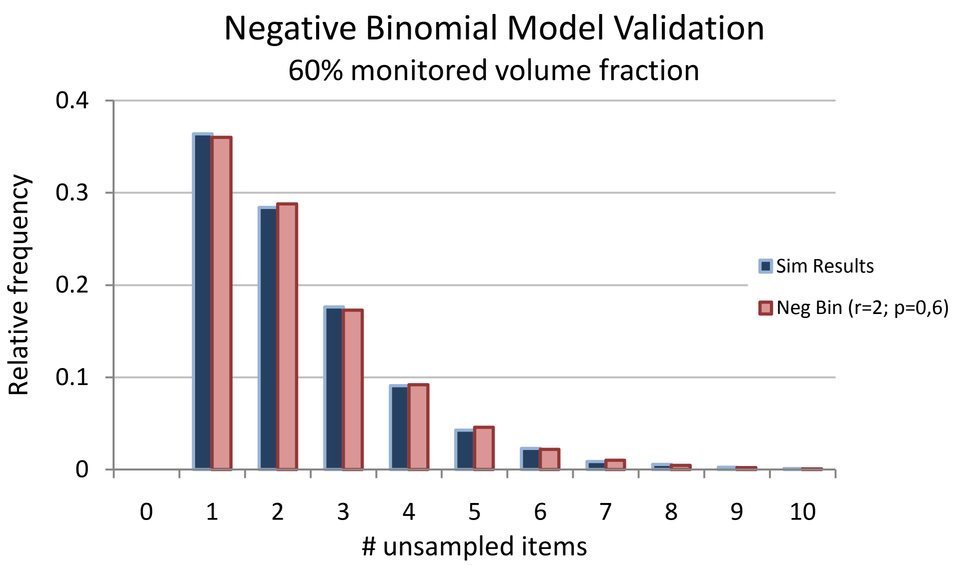


FIGURE 5.35 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: VALIDATION OF THE NEGATIVE BINOMIAL PREDICTION MODEL FOR R=2 AND P=0.6.

When the monitored volume fraction is reduced, the goodness of fit of the negative binomial prediction model is still relevant; however, the prediction error slightly increases. Considering the scenarios in Figure 5.36 and Figure 5.37 the maximum absolute errors increase to 0.97% and 0.92%, respectively; the average absolute errors are 0.38% and 0.35%, respectively.

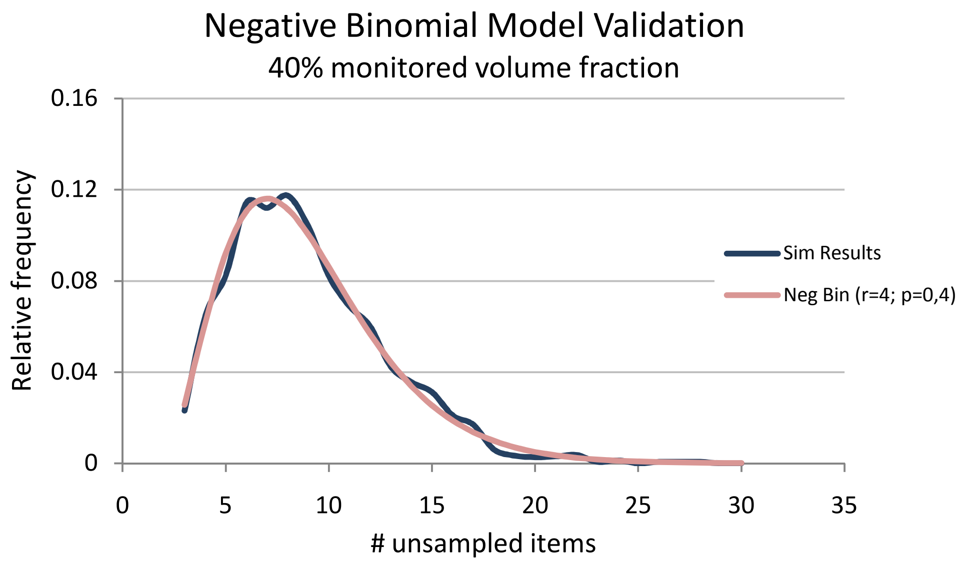


FIGURE 5.36 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: VALIDATION OF THE NEGATIVE BINOMIAL PREDICTION MODEL FOR R=2 AND P=0.4.

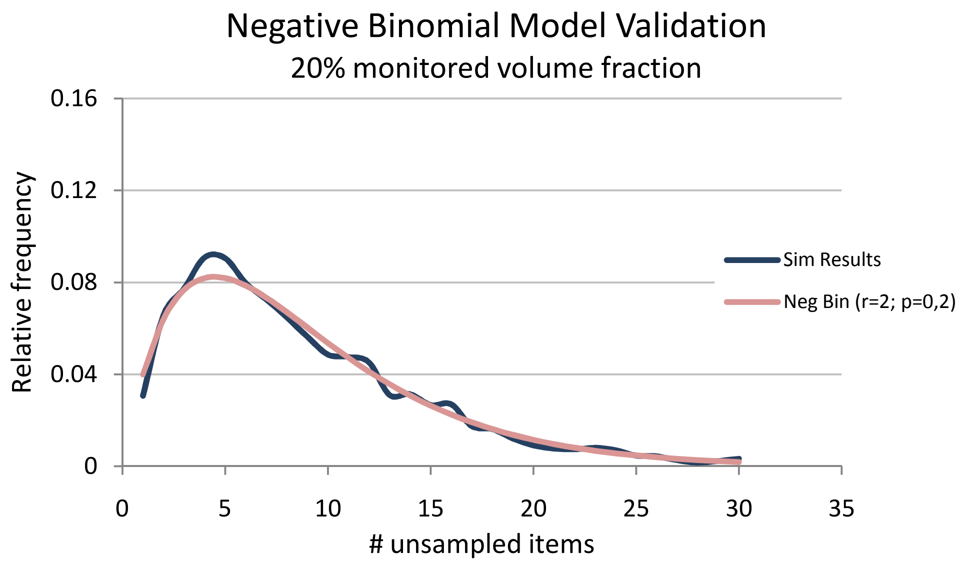


FIGURE 5.37 DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: VALIDATION OF THE NEGATIVE BINOMIAL PREDICTION MODEL FOR R=2 AND P=0.2.

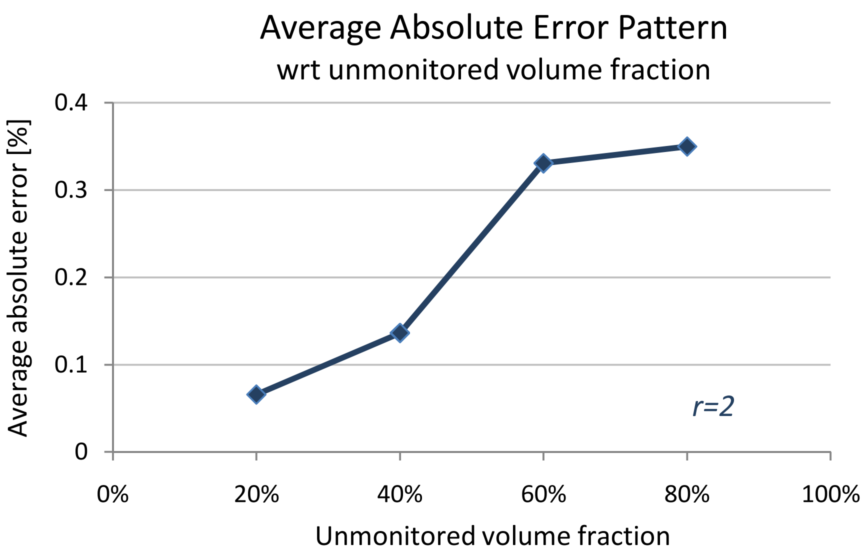


FIGURE 5.38 AVERAGE ABSOLUTE ERROR WITH RESPECT TO THE UNMONITORED VOLUME FRACTION FOR *R*=2.

Figure 5.38 reveals the negative impact of the unmonitored volume fraction on the average absolute prediction error, while the sampling interval is kept equal to 2. The increasing pattern can be explained with consideration of the random nature of the unmonitored flow. In fact, the randomness of the number of unsampled items for the combination flow analysed in this section is entirely due to the presence of the unmonitored flow. The larger its volume is, the larger the variability that characterises the system. In the scenarios simulated, this is also stressed by the fact that the inter-arrival time distributions are exponentially shaped; hence, larger unmonitored volume fraction not only correspond with larger mean inter-arrival time, in comparison with the monitored type, but also larger variability of the inter-arrival time. This introduces higher dispersion, which also determines the presence of numbers of unsampled items much higher than the typical values in the distribution studied in this thesis.

The impact of the sampling interval on the prediction error proves less relevant than the impact of the monitored volume fraction. This is shown in Figure 5.39 where the error pattern can not be clearly defined. However, excluding the second point a slightly increasing trend can be noticed. This could be caused by a cumulative dispersion effect which a larger sampling interval creates. In fact, when sampling intervals greater than 1 are considered, the number of unsampled items between consecutive samples can be considered as the number of unmonitored items between consecutive monitored items times the sampling interval. This is equivalent to summing up identically distributed geometric random variables. The resulting distribution, which is negative binomial, is characterised by a variance greater than the variance of the originating distribution and linearly dependent on the number of added variables, which in this case is the sampling interval. So the greater the sampling interval, the greater the dispersion, more frequent the anomalous values and, finally, greater the prediction error.

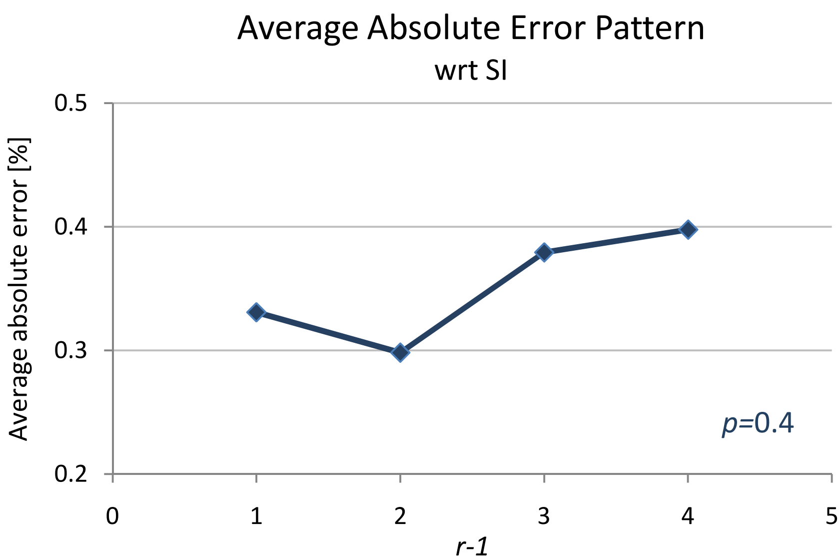


FIGURE 5.39 AVERAGE ABSOLUTE ERROR WITH RESPECT TO THE SAMPLING INTERVAL FOR P=0.4.

5.5.2 Two monitored products case

When two different products are monitored, the resulting distribution for the number of unsampled items between consecutive samples does not have an immediate reference to any of the most common discrete distributions. This might appear to be a simple case, since cross flow products do not interfere with the sampling process and both the product types processed in the station are subject to a deterministic sampling plan. However, the combination, or the sum, of two degenerating discrete distributions, that is, distributions characterised by only one value with frequency equal to one, does not generate a discrete distribution of the same type. For example, the case with two monitored products produced in the same volume fraction with the exact same sampling interval can be misleading. In fact, it would be easy to erroneously consider the distribution of the number of consecutive unsampled items is still a degenerating distribution characterised by a single value domain, the common sampling interval minus one, with frequency equal to one. However, when two products, in the same volume fraction, with different sampling frequencies, for instance equal to 4 and 3 respectively, are considered, it should be immediate to understand that a distribution with a domain consisting of only one integer value is not realistic. First, because the sampling intervals are different; second, because, even if the average sampling interval is evaluated it can result in a non integer value, in this case 3.5, and obviously this does not agree with the discrete nature of the distribution. As a consequence the distribution of the number of unsampled items when two monitored products are present in the sampling station is surely a non-degenerative distribution. On the other hand, as seen before, when variables described by distributions of the same nature are summed, the resulting variable is not necessarily described by the same distribution type (e.g. the sum of geometric distributions is a negative binomial distribution).

The only aspect easy to deduce in this particular context concerns the domain of the distribution. In fact, since the two distributions which generate the final one are characterised by only one value, it should be straightforward that the maximum value of the domain of the final distribution, which can roughly be considered a weighted sum of the two distributions, is given by the sum of the two values in question. This appears even simpler by reflecting on the meaning of the variables summed. Considering a practical example, if the first product type is sampled every third item (two items are skipped and the third one is measured) and the second product type is sampled every fourth item (three items are skipped and the fourth one is sampled), it is easy to agree that in the worst case scenario from a risk quality perspective, a maximum number of five consecutive items can miss the sampling decision. This consideration has been verified by simulation experiments.

Along with scenarios available from previous analyses (as scenario 1 (Table 4.1 (p. 89)) and scenario 11 (Table 5.2 (p. 130)) and scenarios considered for the non-sampling station case (Table 5.4 (p. 142)), new scenarios were simulated for the study of the sampling station case; since the procedures developed were based on logical reasoning, the most of them proved useful for the validation process. The new scenarios differ from each other for both the volume fractions of the three product categories populating the model and the shape of the inter-arrival time distributions. The unmonitored product (Pr3) was not considered for the case analysed in this section; it will be taken into account during the analysis presented in Section 5.5.3. Whenever the lognormal distribution was chosen to model the inter-arrival times, the coefficient of variation was set equal to one in order to allow a statistically consistent comparison between scenarios differing only for the inter-arrival distribution shape. The choice of the log-normal (LN) distribution as an alternative to the exponential distribution for modelling inter-arrival times is due to both its shape, which is suitable for modelling an arrival process, and the possibility to easily control its mean and standard deviation. The scenarios which consider lognormal inter-arrival time distributions are only considered during the investigation on the impact of the time related distributions on the prediction models accuracy. This analysis is extensively presented in Appendix F and summarised in Section 5.5.4. All figures reported in Sections 5.5.2 and 5.5.3 refers to scenarios characterised by exponentially distributed inter-arrival times unless otherwise specified. Table 5.6 lists the new scenarios simulated. The four groups identified by the thicker lines correspond with four different volume fraction ratios between product 1 and product 2. The slightly different ratio between Pr1 and Pr2 volume fractions for scenarios SS5 (and SS7) and SS6 (and SS8) are due to the fact that the fractions are rounded to two decimal figures; the ratio is actually equal to 1.5. For the same reason, the actual ratio between Pr1 and Pr2 volume fractions for scenarios SS1 and SS2 is 3.5. For each ratio, different unmonitored volumes were considered and/or the impact of the inter-arrival time distribution shape was investigated. Scenarios SS10 and SS11 differ from each other for the absolute volumes. In order to avoid repetitiveness, some of the scenarios in Table 5.6 are not included during the results analysis.

The simulation model was run for 10,000 hours with a warm-up period of 1500 hours. 5 replications were conducted for each scenario and data from the model was averaged across the 5 simulation runs. Since different combinations of sampling intervals were considered, the number of samples behind each reported distribution varies; for the smallest sampling intervals (1,1) the sample population size is around

18000, for the largest sampling intervals considered (4,4) the population size reduces to 7000 ca.

TABLE 5.6 SCENARIOS SIMULATED FOR THE ANALYSIS OF THE SAMPLING STATION CASE.

	Pr1		Pr2		Pr3	
Scen.	**IAT dist.**	**Vol. Fr. (%)**	**IAT dist.**	**Vol. Fr. (%)**	**IAT dist.**	**Vol. Fr. (%)**
SS1	Exp.	0.16	Exp.	0.04	Exp.	0.80
SS2	LN	0.16	LN	0.04	LN	0.80
SS3	Exp.	0.46	Exp.	0.23	Exp.	0.31
SS4	Exp.	0.18	Exp.	0.09	Exp.	0.73
SS5	Exp.	0.14	Exp.	0.10	Exp.	0.76
SS6	Exp.	0.34	Exp.	0.22	Exp.	0.44
SS7	LN	0.14	LN	0.10	LN	0.76
SS8	LN	0.34	LN	0.22	LN	0.44
SS9	Exp.	0.25	Exp.	0.25	Exp.	0.50
SS10*	Exp.	0.43	Exp.	0.43	Exp.	0.14
SS11*	Exp.	0.43	Exp.	0.43	Exp.	0.14

* IAT values different for these two scenarios

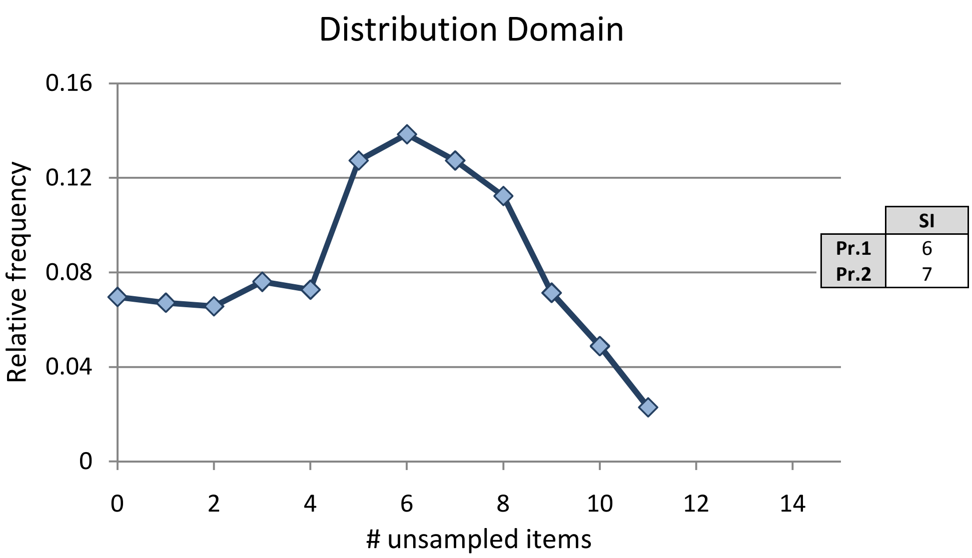


FIGURE 5.40 LIMITED DOMAIN FOR THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN CONSECUTIVE SAMPLES FOR TWO MONITORED PRODUCTS.

For all the combinations of sampling intervals considered, the maximum value of the random variable was exactly as predicted. Figure 5.40 shows the distribution of the number of unsampled items between consecutive samples for a scenario characterised by two monitored products whose volume fraction are 0.6 and 0.4 respectively. The sampling interval is 6 for the first product and 7 for the second product. As expected, the maximum number of unsampled items is 11, which is 6+7-2, where 2 represents the samples, which are not included in the distribution.

If the range of distribution domain is easily predictable, unfortunately, the shape of the distribution is not. In the literature consulted [146, 147] arithmetic operations on continuous and discrete distributions are illustrated, but the case of the sum of two degenerative distributions was not treated. An analytical approach or any aid from the most common distributions does not seem likely, since the shape of the distributions obtained from the simulation experiments does not recall the shape of any well known discrete distribution. In a few cases, the negative binomial distribution fits reasonably well the simulation results. However, there is no theoretical support to this solution. Moreover, the negative binomial distribution does not present a limited domain, even though it usually generates very low values for the upper tail.

Since, an immediate application of a classical distribution was not possible, another basic approach was considered. It consisted of enumerating all the possible sequence combinations for a given sequence length and generating the final distribution as a weighted average of the distribution associated with each combination. Enumerating techniques are treated in Meyer [147] and they seem to provide a solution to all the statistical problems when application of an analytical approach is difficult or impossible. This approach presents a combinatorial nature which, apart from very banal cases, makes it manually inapplicable. Appendix C illustrates a few functions developed. However, owing to the limited specification of the computer used, only a limited range of sampling interval combinations could be analysed in a reasonable time period.

The algorithm developed, based on an enumerative approach, consists of a few steps which will be described in Section 5.5.2.2. Before that, the next section introduces the inputs needed to start the algorithm. These include the number of iterations which the algorithm should be run for, the distribution of the number of items of the second (/first) product between two consecutive items of the first (/second) product and the sampling intervals of the two products.

5.5.2.1 *Inputs description*

The indication of the number of iterations which the algorithm has to be run for, besides representing a fundamental parameter for the correct execution of the algorithm has a further meaning. It is related to the maximum length of the item sequences which will be analysed during the procedure. When the algorithm is applied from the product 1 perspective, the number of product 2 items is increased every time a new iteration starts. Hence, the number of iterations determines the maximum number of product 2 items which will be included in the item sequences analysed. In order to simplify the illustration of the concepts described in this section, a common reference scenario is set for the examples shown. In this scenario, products 1 and 2 are monitored with sampling intervals equal to 4 and 6, respectively; their volume fraction is equal to 0.4 and 0.6, respectively. If the algorithm is first applied from the product 1 perspective, which means, as it will be shown later, that a different number of product 2 items will be progressively considered, a number of iterations equal to 15 means that the algorithm will be applied 16 times on increasingly longer product 2 item sequences. In this case, the product 2 sequence length will range between 0 and 15 items. The actual overall sequence length is obtained by summing up the number of product 1 items present in the sequence. This number does not change throughout the different iterations and it is equal to product 1 sampling interval plus one. For the example considered, the overall sequence length will range between 5 and 20 (Figure 5.41). The mechanisms for the creation of the sequences will be illustrated during the description of the algorithm steps (see Section 5.5.2.2).

In order to give statistical consistency to the item sequence analysis, the number of iterations should be chosen in relation to the sampling intervals analysed. In particular, in order to explore different sequence combinations, it would be advisable considering a maximum number of items for the mobile product, which, in this example, is product 2, ranging between two or possibly three times its sampling interval.

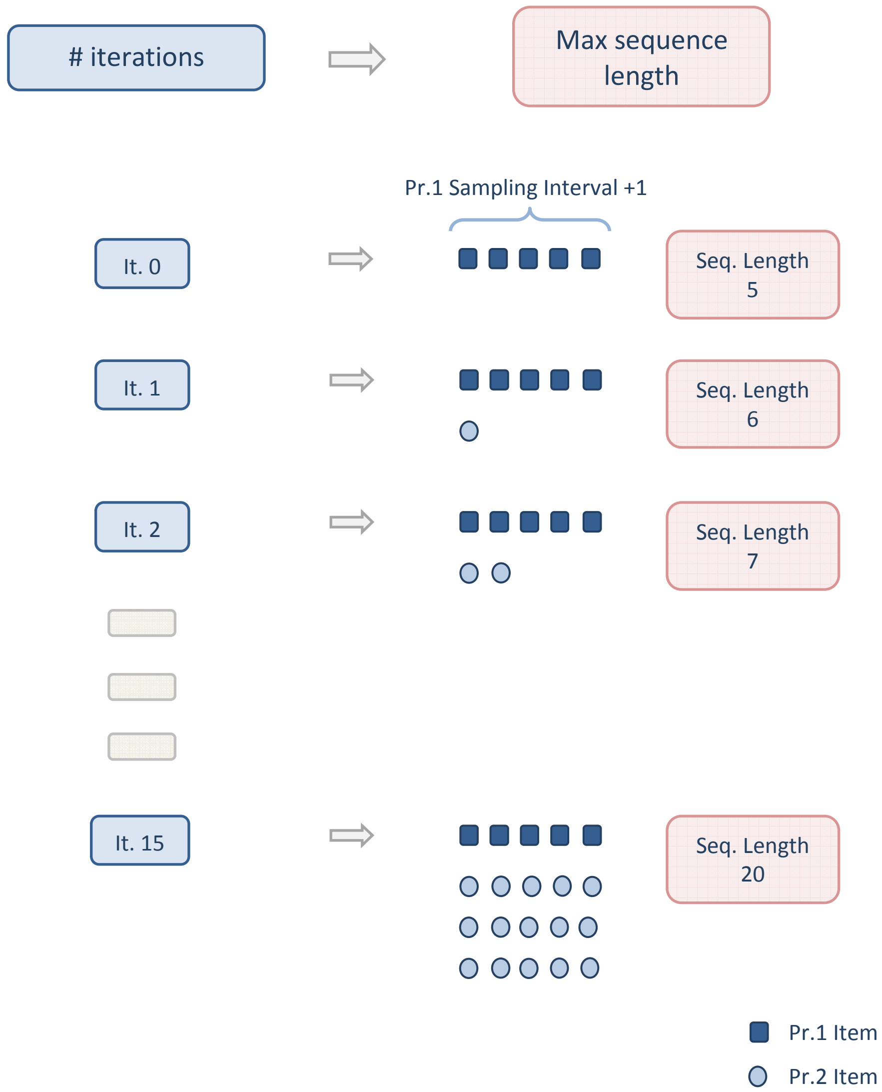


FIGURE 5.41 IMPACT OF THE ITERATION NUMBER ON THE ITEM SEQUENCE LENGTH.

The distribution of the number of product 2 (product 1) items between consecutive product 1 (product 2) items deserves more attention. It is useful for evaluating with which probability items of the same product can be consecutively produced. Hence, it also supports the calculation of the probabilities associated with each item sequence combination generated during the algorithm; the probabilities will serve as weighting factors in the final distribution sum. The distribution of the number of consecutive items of the same product is not immediately available; however, two approaches can be used in order to build it. The simpler approach is only feasible when historical data are available, either coming from a factory database or from simulation results. Based on those data, the distribution can be immediately derived. Another way to calculate the distribution is based on its meaning; it can be defined as the distribution of the random variable describing the number of consecutive failures, in this case the number of consecutive items of the same product, before the first success, that is, the first item of the other product. Assuming that the trials are independent and the probabilities associated with the two events do not change over time, the distribution under investigation can be considered geometric with a proportion equal to the volume fraction of the product which is associable with the successful event. Simulation experiments show that, in most cases, this approximation can be considered good.

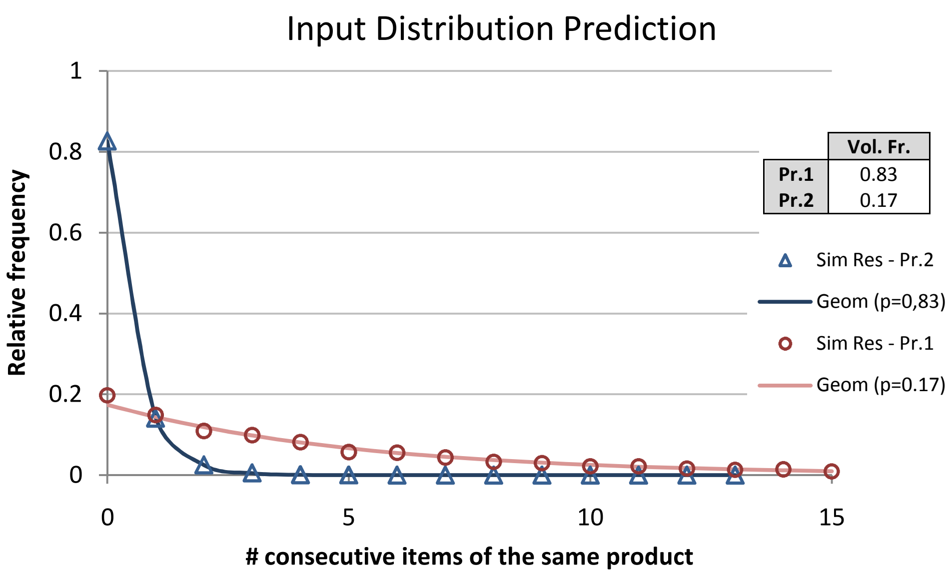


FIGURE 5.42 EFFICACY OF THE PREDICTION MODEL FOR THE DISTRIBUTION OF THE NUMBER OF CONSECUTIVE ITEMS OF THE SAME PRODUCT (SCENARIO 1)

In Figure 5.42, the efficacy of the geometric distribution in predicting the distribution of the number of consecutive items of the same product is shown from the double perspective of product 1 and product 2 for a scenario different from the one set at the beginning of this section (p. 187). The volume fraction of product 1, in this first scenario, is 0.83. The distribution of the inter-arrival time for both the products is exponential. The agreement of both the distributions with the geometric prediction is very good, with maximum absolute errors equal to 0.29% and 0.95% for consecutive product 2 and product 1 items, respectively. The higher error for the consecutive product 1 items distribution is probably due to the fact that being the inter-arrival time of product 2 items characterised by greater variability, the time range between two consecutive product 2 items is definitely less precise than between two consecutive product 1 items. The chi square test provides statistical evidence of the goodness-of-fit of the geometric predictions; the P-values are 0.03 and 0.16 for product 2 and product 1, respectively.

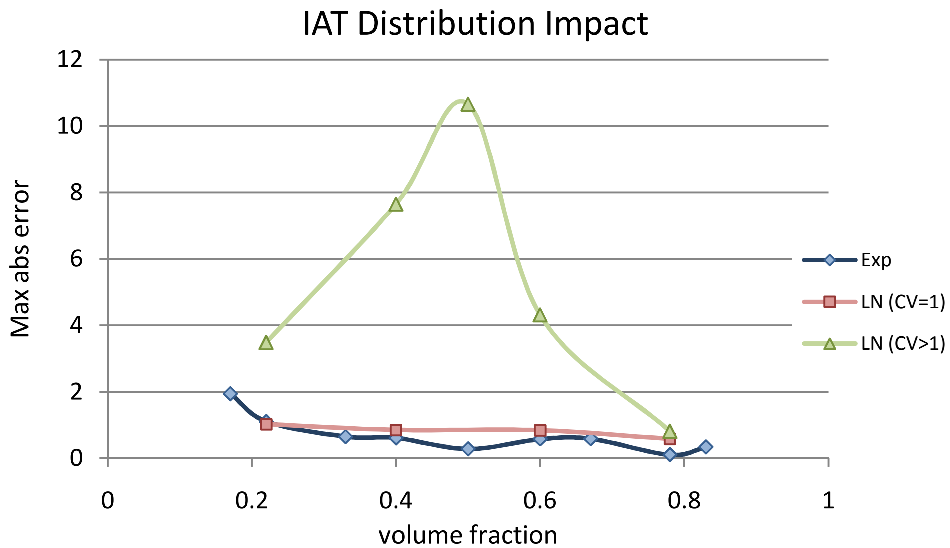


FIGURE 5.43 IMPACT OF THE INTER-ARRIVAL TIME DISTRIBUTION ON THE EFFICACY OF THE GEOMETRIC MODEL PREDICTION FOR THE DISTRIBUTION OF THE NUMBER OF CONSECUTIVE ITEMS OF THE SAME PRODUCT.

The impact of the shape of the inter-arrival time distribution and its variability on the distribution of the number of consecutive items of the same product was investigated. The results obtained, shown in Figure 5.43, reveal that the inter-arrival time distribution has an impact on the goodness-of-fit of the geometric prediction. With respect to the exponential inter-arrival time scenarios, the maximum absolute error slightly increases for the lognormal distribution scenarios, when the coefficient of variation is equal to 1 (Figure 5.43). As Figure 5.43 shows, the variability of the inter-arrival time distribution relevantly impacts the prediction model accuracy; when the log-normal distribution has a CV greater than 1 the pattern of the prediction error presents a peak for central values of the volume fraction.

Figure 5.43 also suggests that the volume fraction, which is the only parameter of the prediction model, does not affect the prediction error when the inter-arrival time variability is low. The error pattern presents a peak in correspondence with 50% volume fraction for the high inter-arrival time variability case. This is presumably due to the fact that, in this case, there is no predominance of one product on the other and the resulting scenario can be considered the most uncertain since it is like the effects of the high variability are doubled. It is worth noting that for the errors reported in Figure 5.43 no distinction has been made between the two products; low volume fractions are usually associated with product 2 and high fractions correspond with product 1. However, given the perfect equivalence of the two products, it is possible to state that similar results would be obtained if the product fractions were switched around. Whenever more than one value was available for a particular volume fraction, the average between the values is reported in the graph.

From the results obtained, the prediction model for the distribution of the number of consecutive items of the same product seems to work reasonably well. It could be refined with consideration regarding the characteristics of the inter-arrival times distribution. Unless otherwise specified, the distributions of the number of consecutive items of the same product obtained from simulation results will be used for the prediction of the distribution of the number of unsampled items between consecutive samples in all the applications proposed later on. This is to avoid any cumulative effect of the prediction error, and, hence, any misleading evaluation of the efficacy of the algorithm developed for the prediction of the distribution of the number of consecutive unsampled items.

5.5.2.2 Algorithm steps

Given that the inputs required have now been defined the algorithm development will now be discussed. The algorithm in effect involves the solution of two problems each of which is associated with one product. The algorithm steps will be described for the product 1 perspective but the same procedure has to be applied to product 2 as well.

Considering product 1 perspective means that product 1 will be treated as a rigid product; its sampling interval univocally determines the number of product 1 items to be considered in all the sequences which will be generated during the algorithm. Every sequence will have as the first and the last item a product 1 item; a number of product 1 items equal to the product one sampling interval minus one will be allocated within the extremes of the sequences generated by the algorithm. For instance, if product 1 sampling interval is four, five product 1 items will be present in any sequence generated by the algorithm as all the sequences will have a product 1 item as the first and the last item and the other three (four minus one) items will be placed in the middle of every sequence.

Each sequence is also characterised by the presence of a quantity of product 2 items which depends on the particular iteration run as one more product 2 item is included when a new iteration starts. The algorithm starts with no product 2 items. This means the algorithm starts with only one sequence, exclusively composed by the fixed number of product 1 items, five in the example previously considered. As a consequence, there is only one permutation for the initial iteration (iteration #0) as the probability that a sequence of five type 1 items can be consecutively processed by a machine is independent of the order with which those items move out. The probability associated with an item sequence generated by the algorithm represents the probability to observe items moving out of a machine in the particular order defined by the sequence. This probability can be calculated as a simple compound probability. In fact, any sequence can be decomposed in partial sequences delimited by product 1 items (product 2 items when product 2 perspective is considered). Depending on the position of product 1 items in the original sequence, the length of the partial sequences can vary. The probability associated with a partial sequence is the frequency, in the distribution of consecutive product 2 items, corresponding with the number of product 2 items in that partial sequence. The probability associated with a sequence is calculated as the product of the probabilities associated with each single partial sequence of consecutive product 2 items. In the example considered, in the initial iteration, the probability associated with the only possible sequence is given by elevating to the power of four the frequency corresponding with zero consecutive product 2 items in the distribution of consecutive

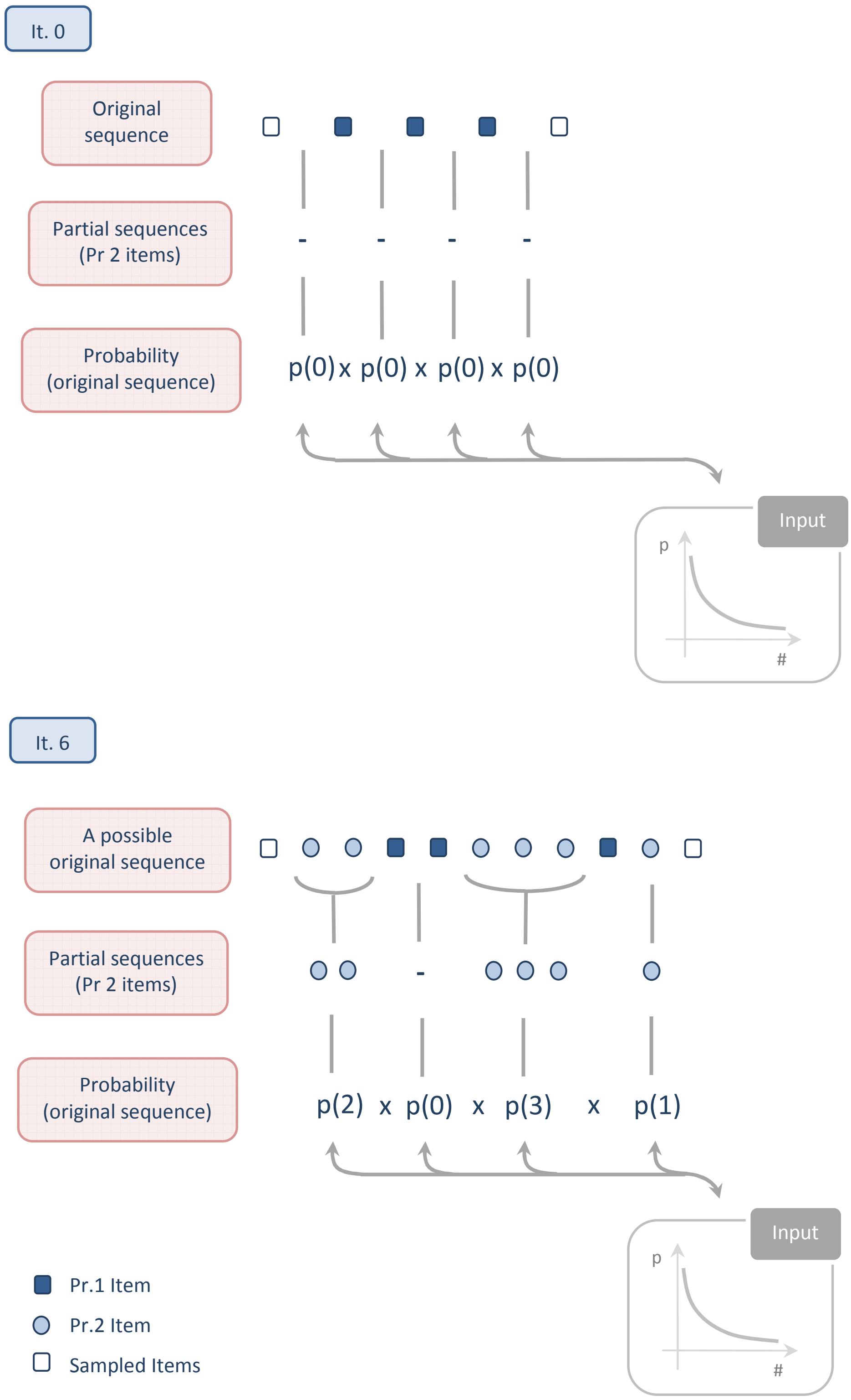


FIGURE 5.44 CALCULATION OF THE PROBABILITY ASSOCIATED WITH AN ITEM SEQUENCE.

product 2 items (Figure 5.44). The elevation to the power of four is the product of the frequency by itself for four times, four being the partial segments limited by product 1 items in the original sequence. In the case considered, each of the four partial segments presents a length equal to zero. Figure 5.44 also shows the calculation of the probability associated with a possible sequence in the fifth iteration. It is worth noting that the hypothesis, made here, of the independence of the length of each partial sequence is reasonable because of the geometric nature of the item departure process. Therefore, the simple product of the probabilities associated with the partial sequences is acceptable.

From each sequence, different distributions of the number of unsampled items between samples can be generated; they are stored along with the corresponding probabilities.

In the second iteration one product 2 item is introduced. This item can occupy four different positions in the sequence of five product 1 items that still do not have any relevant difference from a probability calculus perspective; in fact, no matter where the product 2 item is placed, one partial sequence of one product 2 item and three partial sequences of zero product 2 items will be obtained. However, in this iteration, the position of the product 2 item is meaningful from the perspective of the definition of the distribution of the number of consecutive unsampled items. In fact, the product 2 item introduced might be a sample. This means that the distribution of the number of unsampled items between consecutive samples will change according to the product 2 item position. For each of the four possible item sequences, two distributions can be derived, one considers the product 2 item as a sample, the other one not (Figure 5.45). The two distributions derived will be assigned an equal weight in order to develop the distribution of the number of unsampled items associated with the originating sequence. The sequences, the corresponding probabilities and the calculated distributions will be stored before the next iteration is started.

In general, at each iteration, a further product 2 item is introduced (Figure 5.46). All the possible sequences relevant from a probability perspective are generated. The probability associated with each sequence is calculated. For each sequence all the possible distributions of the number of unsampled items compatible with the product 2 sampling interval are generated (Figure 5.47). This means that at most a number of distributions equal to the product 2 sampling interval minus one will be taken into

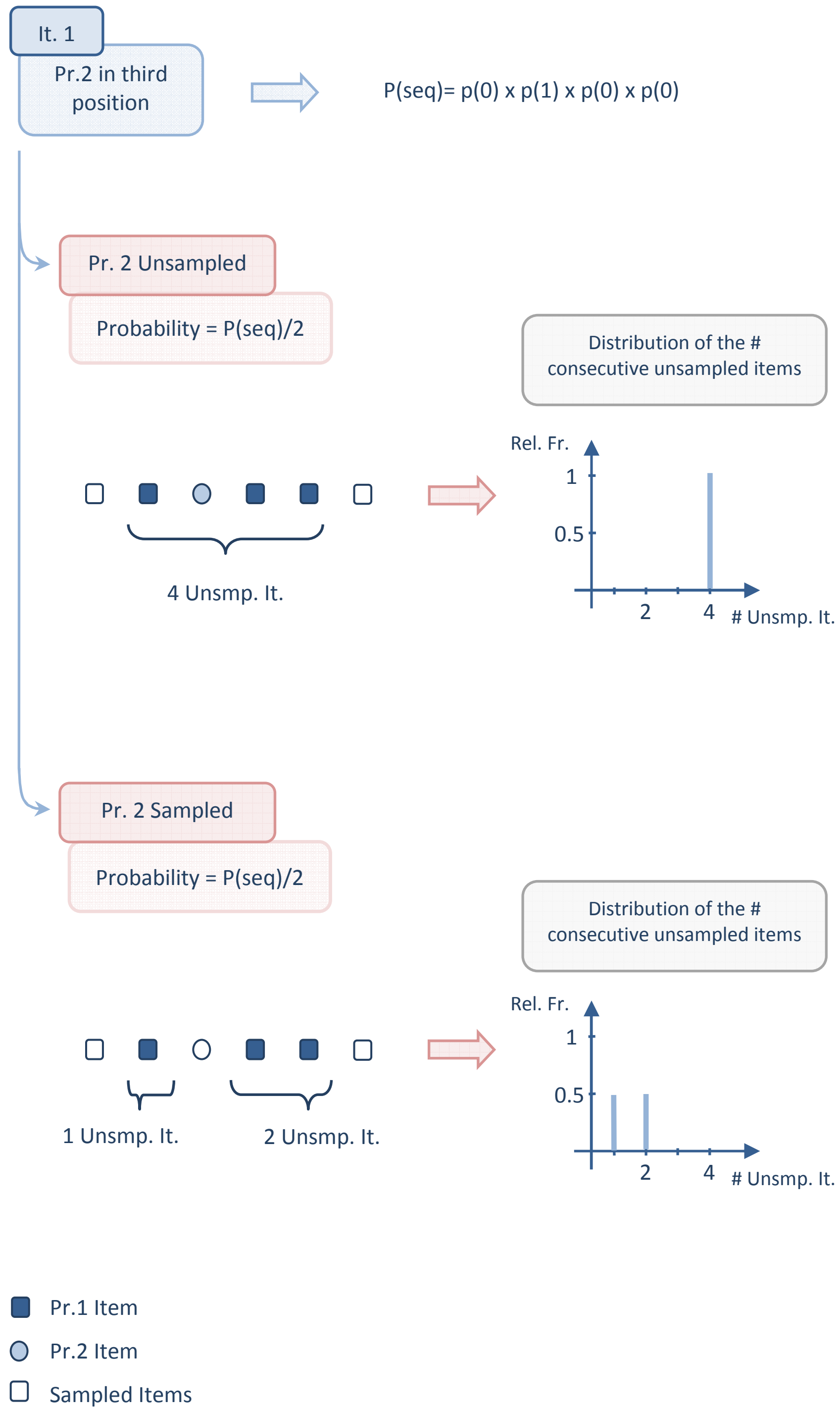


FIGURE 5.45 GENERATION OF TWO DISTRIBUTIONS OF THE NUMBER OF UNSAMPLED ITEMS FROM THE SAME SEQUENCE.

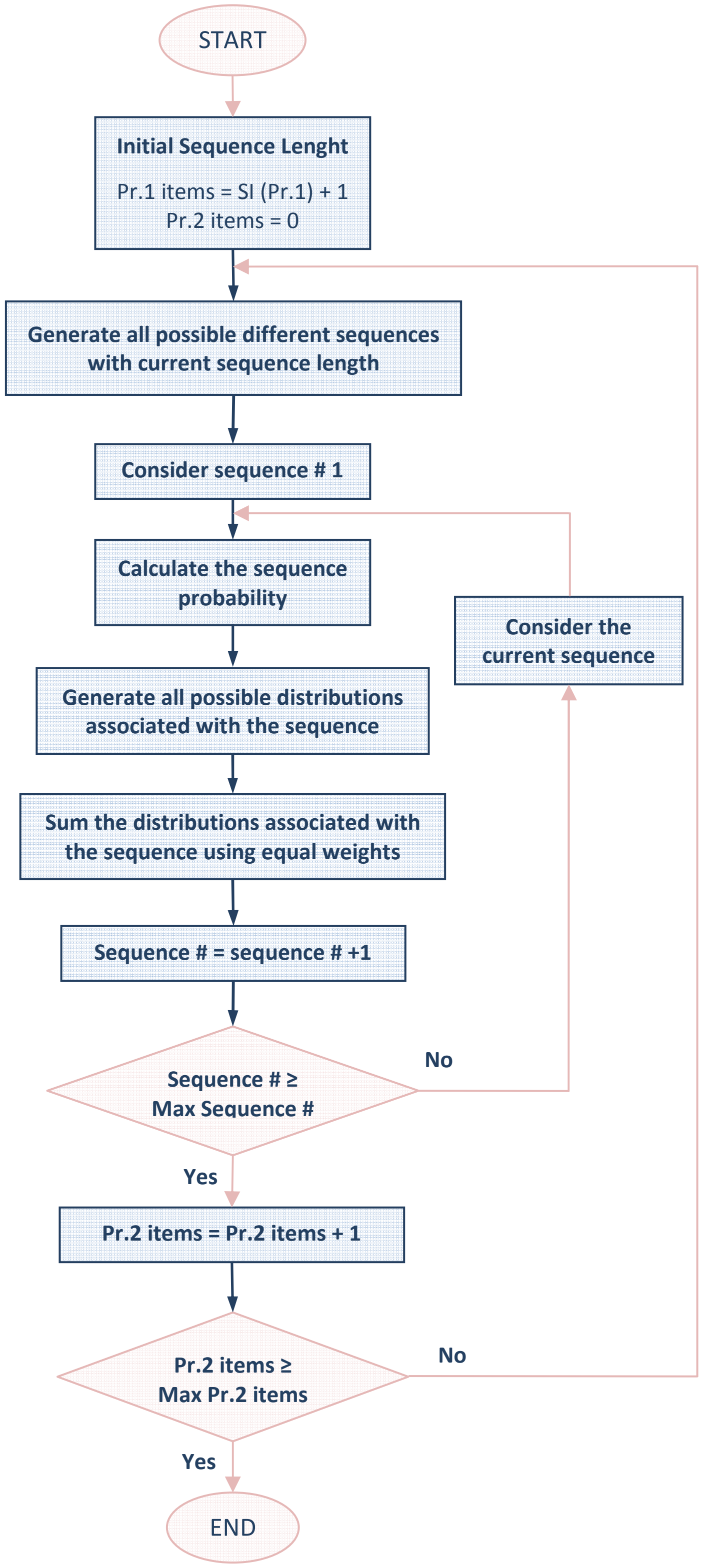


FIGURE 5.46 FIRST PART OF THE PROCEDURE FOR PREDICTING OF THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS (PR.1 PERSPECTIVE).

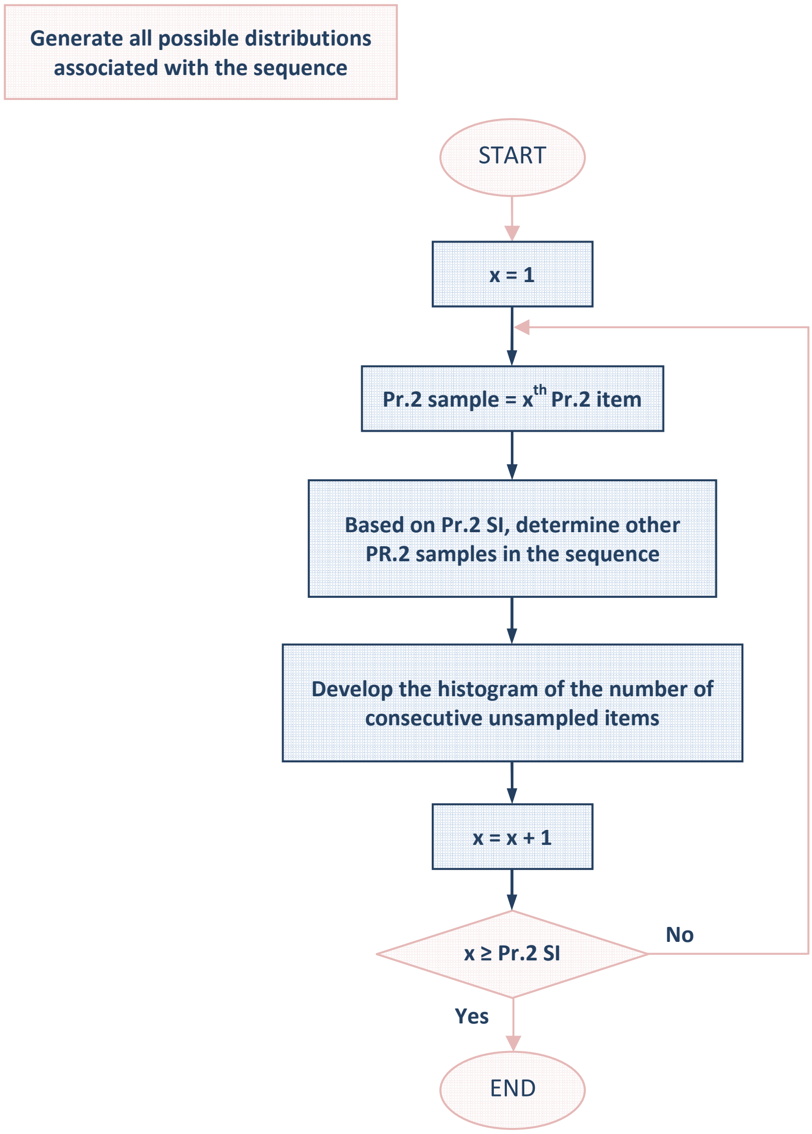


FIGURE 5.47 FLOW CHART REPRESENTATION OF THE GENERATION OF ALL THE DISTRIBUTIONS OF THE NUMBER OF UNSAMPLED ITEMS ASSOCIATED WITH A SEQUENCE.

consideration. In fact, for the way the distributions are generated, a number of distributions greater than the product 2 sampling interval means that identical sequences of sampled and unsampled items would be considered more than once. This situation would not affect the final results since in the moment a greater number of distributions is considered, the associated weights will reduce proportionally; in the end, the total weight associated with a particular distribution would be the same. For each sequence, the derived distributions are summed up using equal weights in order to create the distribution of the number of consecutive unsampled items associated with the originating sequence.

This procedure has to be repeated for the product 2 as well. In fact, so far, the probabilities associated with the sequences of consecutive product 1 items between two consecutive product 2 items have been completely ignored. This can be reasonable only when product 2 represents a very small fraction of the total volume, so that it is very rare that relatively long sequences of consecutive product 2 items can be generated. In other circumstances, ignoring the product 2 perspective would cause a big loss of information and, hence, a not trivial bias of the final distribution.

When the procedure is completed for both the products, a list of distributions with the relative probabilities is available. The final distribution can be obtained from these in different fashions.

In order to consider the perspective of both the products in equal measure, two provisional distributions can be calculated for each product as a sum of each distribution generated by the method weighted by the associated normalised probability. Then the two distributions can be summed up with the same weight or with a weight equal to the volume fraction of the product they are associated with. When the volume fractions are used as weights, this approach will be referred to as the Simple Average (SA) approach hereinafter.

Another way to proceed consists of summing up the distributions regardless of the product they are associated with and weighting them by the corresponding probabilities normalised for both the products. This method seems more rigorous since the relative importance of the distributions associated with the two products is considered. This approach will be referred to as the Weighted Average (WA) approach.

It is worth noting that the two approaches for obtaining the final distribution are theoretically equivalent if the number of iterations run is large enough to produce probabilities whose sum is very close to one for both the products. This means that the algorithm is ignoring only the most improbable item sequences. Generally the WA approach provided better results than the simple average approach. A good example of the better efficacy of the WA approach rather than the SA approach is shown in Figure 5.48. The WA approach seems to make more sense when the sum of the probabilities relative to the enumerated sequences is much less than one. This happens either for calculation time saving reasons or for a reasonable accuracy level reached by a few number of iterations.

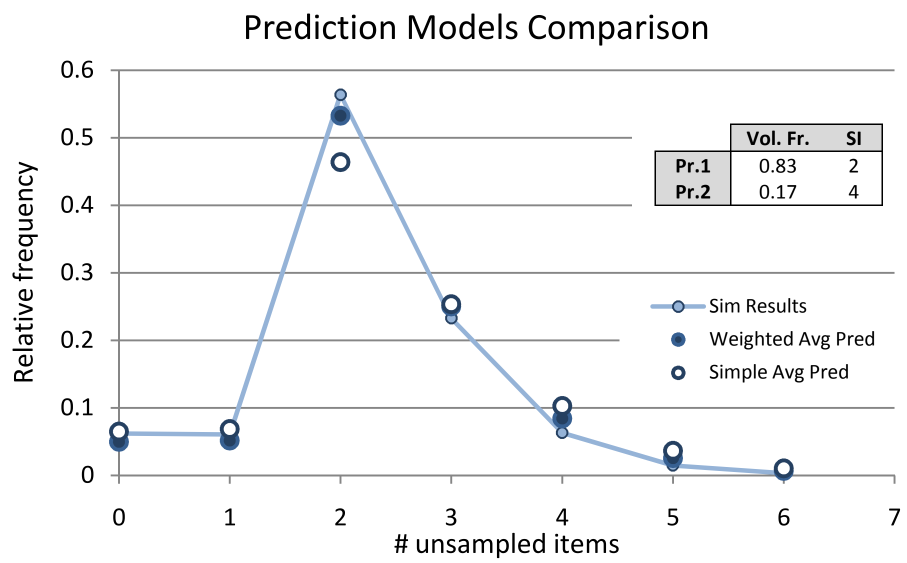


FIGURE 5.48 DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN CONSECUTIVE SAMPLES: COMPARISON BETWEEN THE SIMPLE AND THE WEIGHTED AVERAGE MODEL (SCENARIO 1).

Apart from the combination of the distributions relative to the two product perspective, which was performed using functions developed in Excel, the rest of the procedure was translated into functions developed in Matlab (Appendix C.3).

In the scenario relative to Figure 5.48, product 1 represents 83% of the total volume, with exponentially shaped inter-arrival time distributions for both the products, and the sampling intervals equal to 3 and 5 for the two products, respectively. 15 iterations were run for each product so that 32 distributions were available for the averaging process. The 16 distributions associated with product 1 represent 99.99% of the possible combinations, which means the perspective of product 1 has been fully explored. However, the 16 distributions associated with product 2 constitute only 33% of the possible combinations. This happens because product 2 is less common and less frequently sampled than product 1, and the first distributions generated by the method, which consists of a very small number of product 1 items, represents, in this particular scenario, very improbable situations. In this case, the WA approach makes more sense than the SA approach. Considering a greater number of iterations for product 2, so that a set of combinations with higher total probabilities was available, would be the best way to proceed. That would certainly allow a reduction of the maximum absolute prediction error, which is 3.12% for the WA approach and 10.00% for the simple average approach. The maximum cumulative absolute error is 5.23% for the best performing approach (WA approach). Given the relatively poor specifications of the computer used for the calculations, a greater number of iterations for product 2 was not possible. The computer used has a CPU frequency equal to 2.4 GHz and a 2 Gb RAM. For the limited specifications of the computer available, variations to the algorithm, so that considerations on the sequence probabilities could be included when forming and selecting the sequences, were considered not practical. The advantage of introducing probability considerations during the selection of the sequences to consider for the generation of the distribution could be an interesting element to investigate in the future. It is worth noting that the modifications to the algorithm needed to include probability considerations are minor; in fact, the probability associated with a sequence is calculated prior to the generation of the distributions. This means that a selection of the sequences with the highest probabilities can be made before proceeding with the algorithm. With the current calculator availability, a major concern consisted of the fact that the selected sequences could be very long, so that the computation time would be prohibitive. More powerful calculators would solve this issue.

TABLE 5.7 SAMPLING INTERVALS USED FOR THE MODEL VALIDATION.

	Sampling Intervals								
Pr. 1	2	2	3	3	2	4	3	4	4
Pr. 2	2	3	2	3	4	2	4	3	4

The validation of the prediction model developed was carried out against simulation results. Different scenarios, mostly included in Table 5.6 (p. 185), were considered so that the robustness of the model could be tested. Nine different sampling intervals (Table 5.7) for each volume fraction combination for the two products were simulated.

Recourse to the Pearson's chi square test to formally investigate the goodness-of-fit the WA/SA prediction models was deemed not suitable in this case. Indeed due to the nature of the prediction model, the evaluation of the degrees of freedom to be adopted for the calculation of the test statistics proved prohibitive. Establishing the number of parameters required by the model is not straightforward and, as a consequence, the correction to the number of degrees of freedom can not be performed. Considering that a non-parametric version of the Pearson's test is theoretically possible, the chi square test was applied ignoring the correction for the number of parameters; however, the results obtained revealed a general lack of goodness-of-fit of the predictions for the different scenarios considered. This is due to the non suitability of this particular test. The runs test, also called Wald-Wolfowitz test, was also applied. This test evaluates the goodness-of-fit of a model to an actual distribution by assessing the randomness of the runs of positive and negative prediction errors. Since it is based on signs and not distances, this test is complementary to the chi square test, which, on the contrary, ignores the prediction error signs [148]. Due to the limited and narrow domain of the distributions, the application of the runs test to the predictions developed using the SA/WA approach proves not extremely significant and delivers quite predictable results. For all the scenarios analysed, high P-values were obtained, which means that there is no statistical evidence that the SA/WA models are not suitable for predicting the distributions.

The prediction model proved to be quite accurate, as shown, for example, in Figure 5.49. The shape of the distribution is very well predicted and the prediction errors are reasonably low; the average absolute error calculated on the 5 points of the distribution, is 1.47% and the cumulative error range results 3.68%. It is interesting to note that the peak of the distribution is obtained at the sampling interval (minus one) of product 1, which represents most of the production volume. The case reported in Figure 5.49 represents one of the best cases in terms of prediction accuracy; in general, very good results are also obtained for the other scenarios investigated.

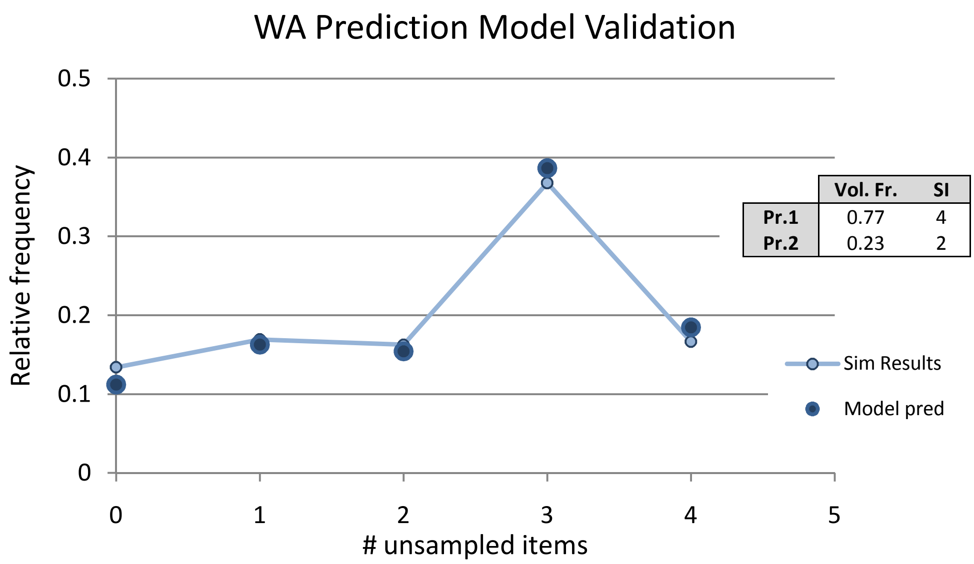


FIGURE 5.49 PREDICTION MODEL VALIDATION: COMPARISON WITH THE SIMULATION RESULTS (SCENARIO SS1).

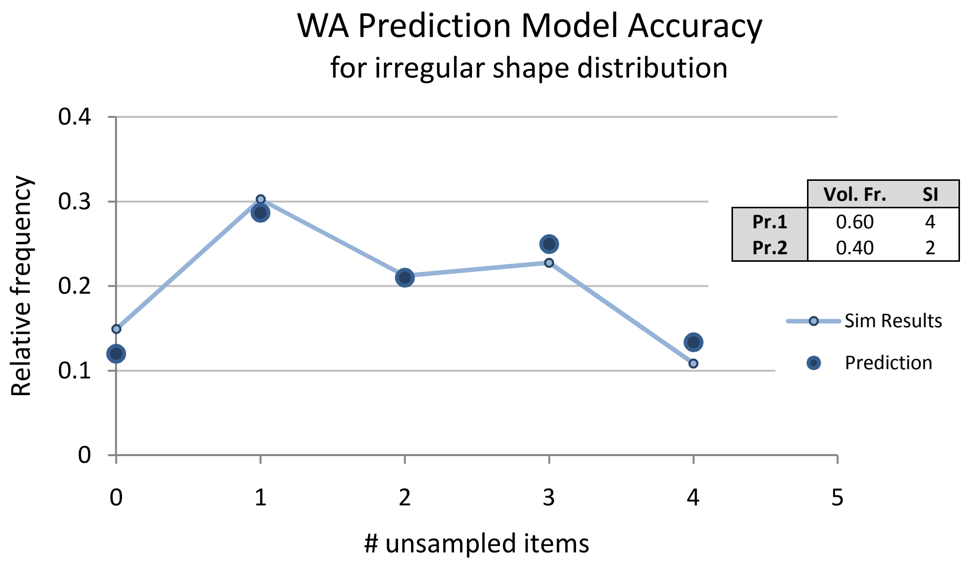


FIGURE 5.50 PREDICTION OF A DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS WITH AN IRREGULAR SHAPE (SCENARIO SS5).

The prediction model works well even when the shape of the distribution is not very regular, as happens, for example, with the distribution represented in Figure 5.50. In this case, the almost balanced production volumes of the two products bestow both the sampling intervals a fundamental role in the shape of the distributions; indeed, two peaks appear in correspondence of the sampling intervals minus one of the two products.

The patterns of both the average absolute errors and the cumulative error ranges were analysed with respect to the sampled fraction, which contains information about the sampling intervals and the volume fractions of the two products.

Despite the irregular pattern shown in Figure 5.51, the average absolute error seems to slightly increase with the sampled fraction. This is reasonable since, keeping the different product volume fractions constant, the smaller the sampled fraction, the greater the sampling intervals. As a consequence, the distribution of the number of unsampled items will have a wider domain and the absolute error will be presumably spread across it.

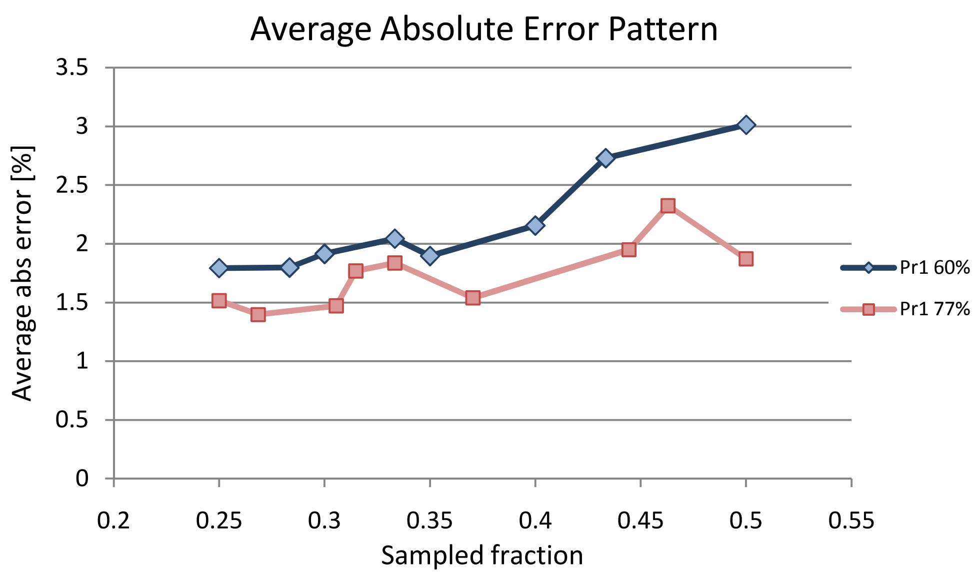


FIGURE 5.51 IMPACT OF THE SAMPLED FRACTION ON THE AVERAGE ABSOLUTE ERROR FOR TWO DIFFERENT PRODUCT 1 VOLUME FRACTIONS (SCENARIO SS5 VS. SCENARIO SS1).

Figure 5.51 also offers another relevant consideration; the product volume fractions evidently affect the prediction accuracy. Bringing product 1 volume fraction from 60% up to 77% reduces the average absolute error by approximately 0.5%, on average. The reason for this is not immediately clear; however, the better results for the higher product 1 volume fraction can be due to the fact that the when the volume fraction is strongly unbalanced, the scenario investigated proves closer to a scenario characterised by only one monitored product, that is a deterministic scenario. The closeness to a regular case could mean less anomalies and, hence, a reduced prediction error.

The two lines in Figure 5.51 refer to the results obtained using the same set of sampling intervals (Table 5.7 (p. 201)). The combination of these intervals with different volume fractions of the two products generates different sampled fractions and, for this reason, sometimes the points of the two lines in Figure 5.51 do not have the same abscissa.

The pattern of the cumulative error range with respect to the sampled fraction, shown in Figure 5.52, proves even more irregular than the one characterising the average absolute error. However, a decreasing trend is still detectable. Even though it might seem incompatible with the average error trend, the pattern can be justified by the way the cumulative error is built. Low sampled fractions generally mean high sampling intervals, that is, a distribution domain consisting of a greater number of values. With an absolute average error that does not relevantly change with the sampling intervals, a greater number of points means a higher chance to increase the cumulative error, in particular when systematic errors affect the prediction model. The impact of the product 1 volume fraction is presumably a consequence of the reduced average absolute error.

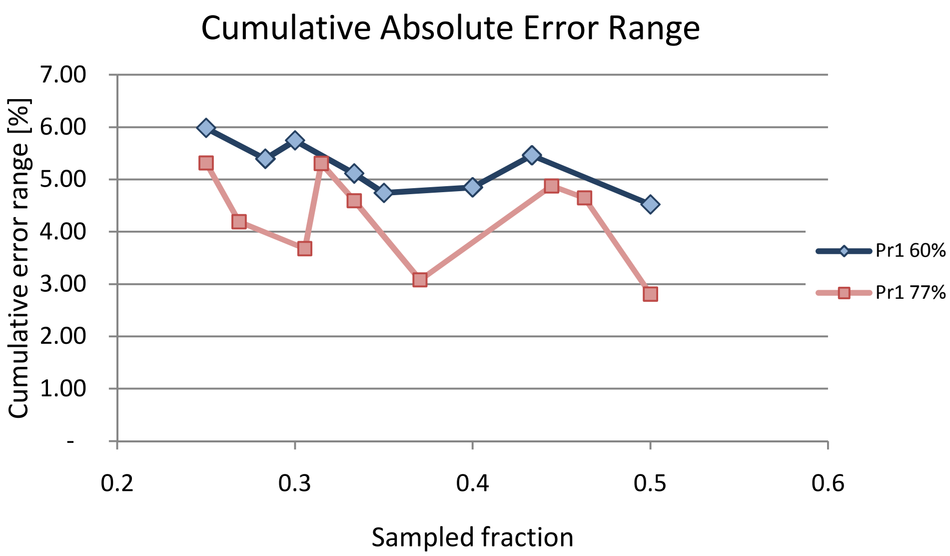


FIGURE 5.52 IMPACT OF THE SAMPLED FRACTION ON THE CUMULATIVE ABSOLUTE ERROR RANGE FOR TWO DIFFERENT PRODUCT 1 VOLUME FRACTIONS.

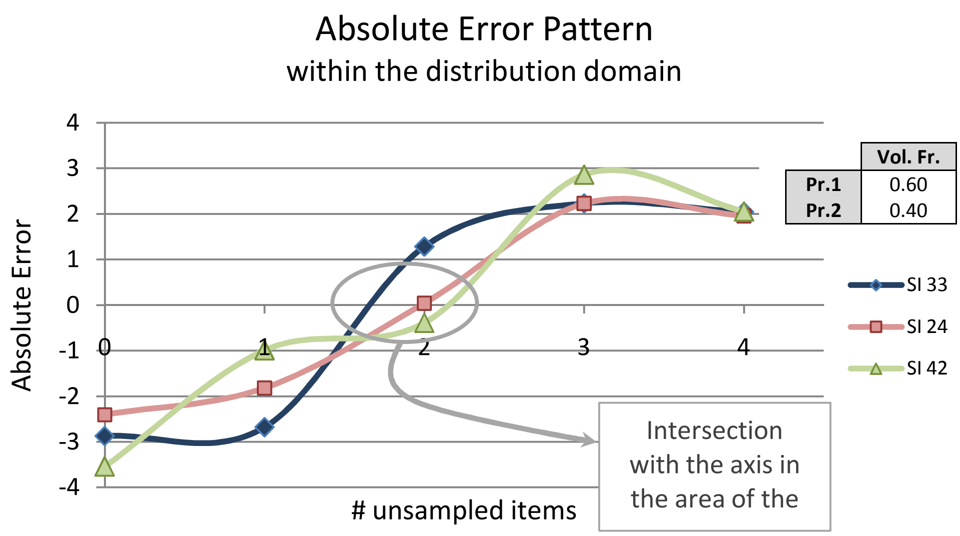


FIGURE 5.53 SYSTEMATIC PATTERN OF THE ABSOLUTE ERROR (SCENARIO SS5).

The analysis of the different results obtained and the conclusions drawn for the cumulative error range, lead to highlight the presence of a systematic pattern in the prediction errors. The predicted distribution always tends to underestimate the frequencies of the lower tail and overestimates the frequencies of the upper tail, generally crossing the simulation results distribution close to its peak (Figure 5.53). The reason for that is probably due to the fact that only very small sampling intervals have been analysed; probably, for larger intervals, the error will spread across the wider distribution domain in a random fashion. Moreover, the fact that only part of all the possible sequences are explored compromises the accuracy of the model, as the results obtained from comparing the SA and the WA approach suggests.

5.5.2.3 Compound sampling intervals

The sampling interval for a product might not be described by an integer, but rather by a set of integers, that is, a cyclic sequence of sampling intervals might be applied. For instance, a product might be sampled following the sequence [1, 2], that is two following items are sampled and one is skipped (and so forth), when that product needs to be sampled on average every 1.5 items. For these scenarios, the algorithm developed can still be used to predict the distribution of the number of consecutive unsampled items.

However, a modification is needed; it is still based on an enumerative approach. It consists of analysing separately all the combinations of the sampling interval values which create the sampling interval sequence. The results obtained combine to create the final distribution. This approach was tested for scenario 11 (Table 5.2 (p. 130)), where the product 1 is subject to a sampling interval described by the sequence [2, 3]. It is obvious that the average sampling interval is 2.5; however, this value can not be directly fed into the algorithm, since only integers are accepted. Product 2 presents a sampling interval equal to 3.

The original problem of predicting the distribution of the number of consecutive unsampled items for sampling intervals equal to ([2,3],3) was split into two different problems. The scenarios derived represent the two possible combinations of the integer sampling intervals, which are (2,3) and (3,3). The distributions obtained from the two scenarios have been averaged to generate the final result (Figure 5.54). The cumulative absolute error range is in this case equal to 3.31%.

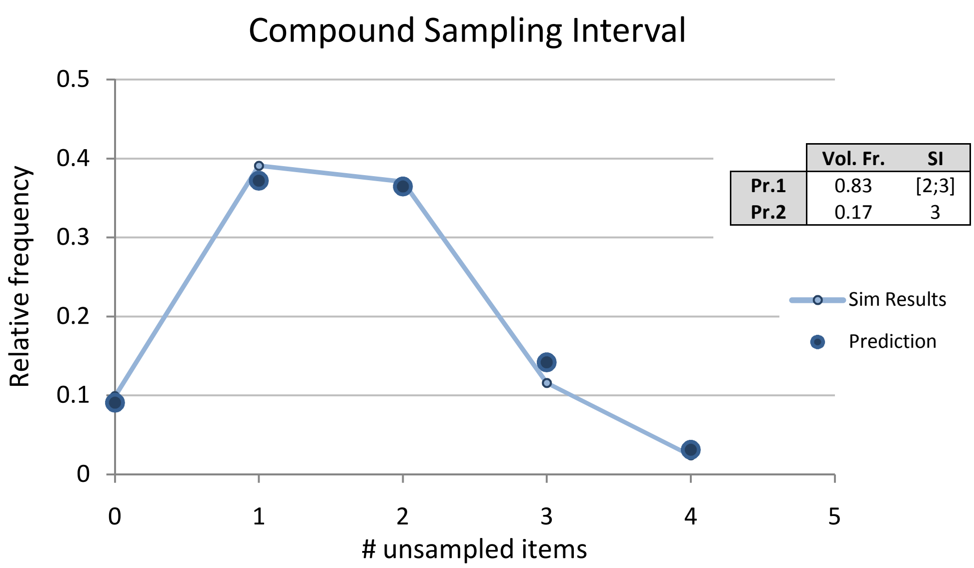


FIGURE 5.54 PREDICTION OF THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN CONSECUTIVE SAMPLES FOR COMPOUND SAMPLING INTERVALS (SCENARIO 11).

5.5.2.4 *Impact of errors in input*

The impact of the use of the theoretical distribution of the number of consecutive items of the same product was also investigated in order to determine the relevancy of the error accumulation effect. With the aim of stressing the eventual negative impact of a poor initial prediction, the analysis was performed on the scenario which represented one of the worst predictions of the distribution of the number of consecutive items of the same product (scenario SS5). This consisted of a volume fraction for product 1 equal to 0.6%; the maximum absolute prediction errors were equal to 4.32% and 7.65% for the distributions relative to product 1 and 2, respectively. The distribution of the number of unsampled items between consecutive samples obtained using the theoretical, geometric, frequencies is plotted in Figure 5.55, where a visual comparison can be made with the distribution obtained from the simulation results and the one obtained by the approach developed using the actual frequencies.

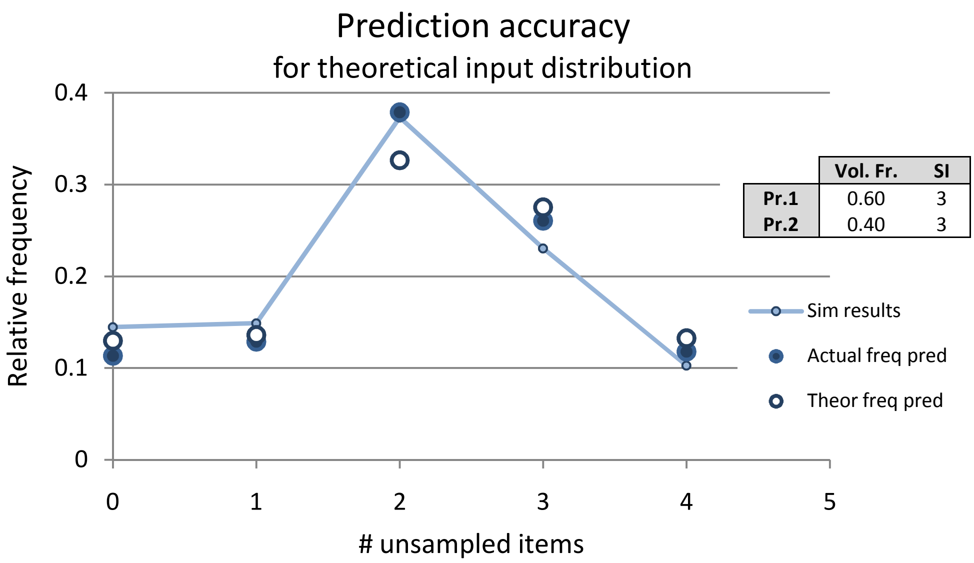


FIGURE 5.55 ACCURACY OF THE PREDICTION MODELS FOR THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS BETWEEN CONSECUTIVE SAMPLES (SCENARIO SS5).

The cumulative prediction error range tends from 5.09% for the actual frequencies up to 7.47% for the theoretical frequencies. The distribution generated using the predicted frequencies in input outperforms the one obtained with the actual frequencies for the lower tail, but presents higher errors for the upper tail. This happens because the theoretical distribution of the number of consecutive items of the same product tends to advantage the higher values, and provides a lower estimation of the frequencies associated with very small sequences of consecutive items of the same product, such as zero or one. This paradoxically causes an increase of the frequencies of the distribution of the number of consecutive unsampled items associated with small values, usually smaller than the modal value and, as a consequence, an increase of the frequencies associated with the larger values of the distribution. In fact, the partial sequences, generated during the application of the approach developed, which contain zero consecutive unsampled items, usually contain a very long sequence of consecutive unsampled items as well. If the theoretical distribution of the number of consecutive items of the same product does not penalise the higher values, as the simulated distribution does, some benefits are paradoxically provided to the smaller values in the final distribution, since what counts in the calculation of a compound probability (as the one associated with the partial distribution is) is not the highest probability of the partial events but the smallest probability. Obviously, as a result the values in the centre of the domain are penalised; hence, the very poor prediction of the peak of the distribution.

Independently of the dynamics behind the nature of the impact of the theoretical input distribution on the prediction of the distribution of the number of unsampled items between consecutive samples, it is interesting to notice that in response to an average error equal to 6.25% introduced as an input, the algorithm produced an increase of the cumulative prediction error range equal to 2.38%. That means that the error in input, even if it has an impact on the prediction accuracy, is not linearly transmitted to the final results. The fact that the input error doesn't linearly propagate through the different stages of the algorithm and it is actually reduced gives confidence in the possibility of using the predicted distribution of the number of consecutive items of the same product when the actual one is not available. This won't cause a relevant loss of accuracy.

5.5.3 Two monitored products + unmonitored flow case

The presence of a cross flow and the combination of two monitored products determine relevant complexities in the analysis of this last case. In fact, unlike what happens for a non-sampling station where the introduction of an unmonitored product flow has the mere impact of scaling the x-axis, in the sampling station, the effect of the introduction of an unmonitored product flow can not be confined to a variation of the axes scale. Even when the simplest scenario, which includes the presence of one-monitored product along with the non-monitored flow, is considered, the effect of the non-monitored flow turns a degenerative probability distribution, consisting of only one value with probability equal to 1, into a non-limited domain probability distribution. In the case when two products are monitored, the original distribution is not so trivial. It has a limited domain and almost predictable shape, with peaks in correspondence with the deterministic sampling intervals. However, the distribution still presents difficulties in deriving predictable frequencies for the different values and its domain obviously contains more than one value.

When the unmonitored flow is combined with two monitored products, an approximation can be made and the approach used for the simpler case of one monitored product can still be followed. Since the dynamics of the item departure process are the same, independent of the particular product flow combination, it is possible to use the negative binomial distribution for an approximated prediction of the distribution of the number of unsampled items between consecutive samples. The average sampling interval, rather than the two distinct sampling intervals, can be used as the parameter r of the negative binomial distribution. The parameter p would be the volume fraction corresponding to the two combined monitored products. The approximation derives from the fact that the hypotheses underlying the negative binomial distribution are not fully respected. This is because a mean value is used as a deterministic value and the pattern of the distribution of the number of consecutive unsampled monitored items is completely ignored. It is worth noting that the use of a non-integer value as a r parameter is not a limiting condition for the procedure as the negative binomial distribution is still applicable for any real value of r. An example of the application of this approximated prediction method, which shows different level of prediction accuracy, is provided in Figure 5.56 and Figure 5.57.`

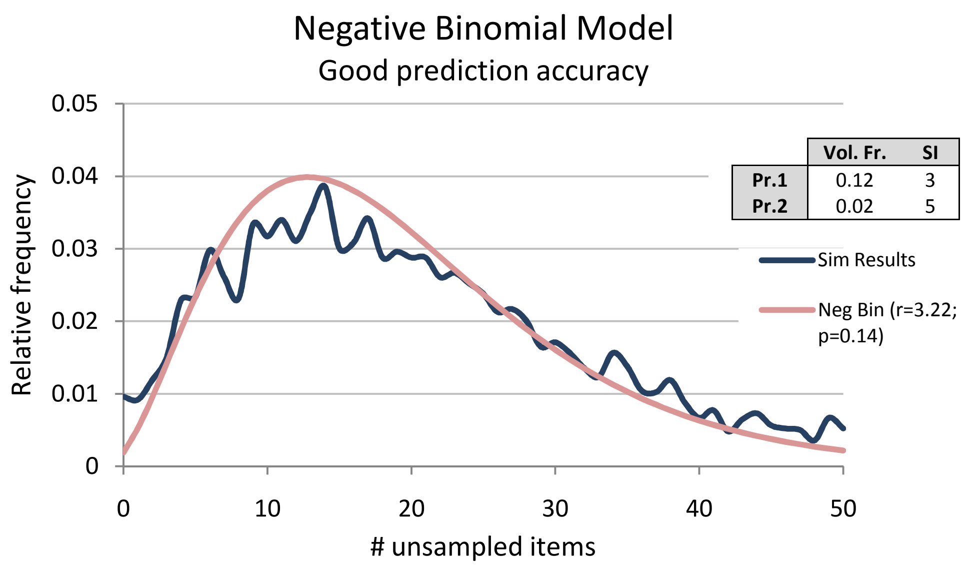


FIGURE 5.56 APPROXIMATED PREDICTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: COMPARISON WITH THE SIMULATION RESULTS (SCENARIO 1).

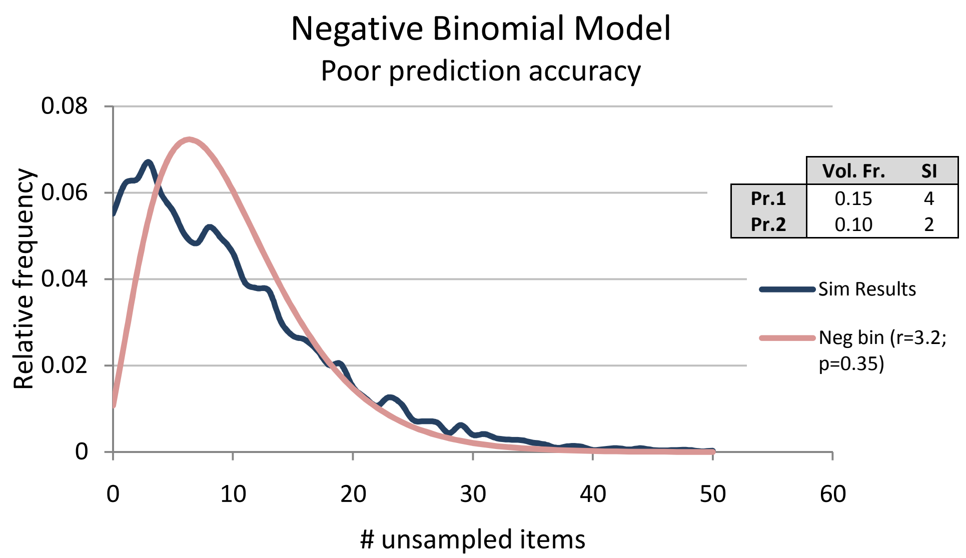


FIGURE 5.57 APPROXIMATED PREDICTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS: COMPARISON WITH THE SIMULATION RESULTS (SCENARIO SS5).

The larger unmonitored volume fraction for the scenario in Figure 5.56 might be the reason why the negative binomial distribution fits the simulation results better than in the case of Figure 5.57. In fact, in presence of a greater unmonitored fraction, it is presumable that the shape of the distribution of the number of consecutive unsampled monitored items has a smaller impact on the final distribution. This means that the average sampling frequency can be used as a reasonably good approximation for the entire distribution. For the two scenarios illustrated, the cumulative error ranges are 8.40% and 16.22%, respectively. The runs test indicates a relatively poor fit for both the scenarios; the P-values are 0.14 and 0.0124 for the distributions in Figure 5.56 and Figure 5.57, respectively. The distributions of the inter-arrival time for all the products are exponential.

Observing the results obtained, the negative binomial distribution represents a good starting point; however, improvements are possible. Relevant considerations for the development of an improved prediction model are:

- the distribution of the number of unsampled monitored items is characterised by a limited domain and it is predictable;
- the negative binomial distribution works reasonably well for the prediction of the number of consecutive unsampled items distribution in presence of an unmonitored cross flow and one product with deterministic integer sampling interval.

A prediction approach which takes into account both these considerations should provide better results.

An immediate way to combine the two elements consists of developing partial distributions for each value of the domain of the distribution of the number of unsampled items between consecutive samples associated with the corresponding no-cross flow case. Since considering a single value of the domain of the distribution of the consecutive unsampled monitored items is equivalent to setting a global deterministic sampling interval for the monitored flow, the partial distributions associated with that single value theoretically should follow a negative binomial distribution. The parameter p will be common to all the partial distributions and equal to the monitored volume fraction. The other parameter, *r*, will be set equal to each value of the distribution of the consecutive unsampled monitored items variable plus one; it will be different for each iteration of the algorithm, that is given by the width of the domain of the originating distribution. The correction used for the parameter *r* is due to the fact that the original distribution considers the number of unsampled monitored items only whereas, in the approach developed, the number of successful trials of the negative binomial distribution also includes the sample. Once the partial distributions are developed they are shifted by the associated value of the originating distribution, unless the definition of the negative distribution as given in [34] is used. The shifted distributions are then summed up by using, as weighting factors, the frequencies associated with the corresponding *r* values in the distribution of the consecutive unsampled monitored items. The last approach/model introduced in this section will be referred to as Combined Negative Binomial (CNB) approach/model hereinafter. The denomination of this approach originates from the nature of the model which combines the negative binomial distribution with the distribution developed using an enumerative approach for the case of two monitored products. At all effects, the distribution resulting from the combined model can be considered a compound distribution.

5.5.3.1 Validation

The CNB approach seems to generate reasonably good results, in particular for the scenarios characterised by a very low monitored fraction. Figure 5.58 and Figure 5.60 show the results obtained for the two scenarios previously considered in the validation of the negative binomial model and respectively represented in Figure 5.56 and Figure 5.57. Figure 5.59 refers to a scenario characterised by the same monitored volume fraction as the one in Figure 5.58 but a reduced sampling interval (scenario 11). The representation of the results obtained for three scenarios is due to the intention of showing the differences in the CNB model prediction performances with respect to the monitored volume fraction and the sampling interval.

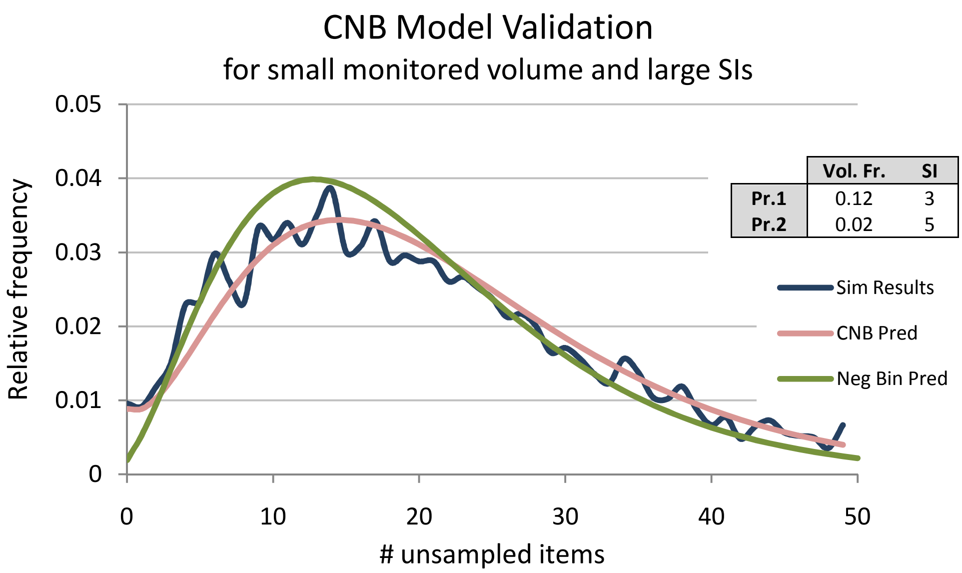


FIGURE 5.58 COMPARISON WITH THE SIMULATION RESULTS FOR SCENARIO 1.

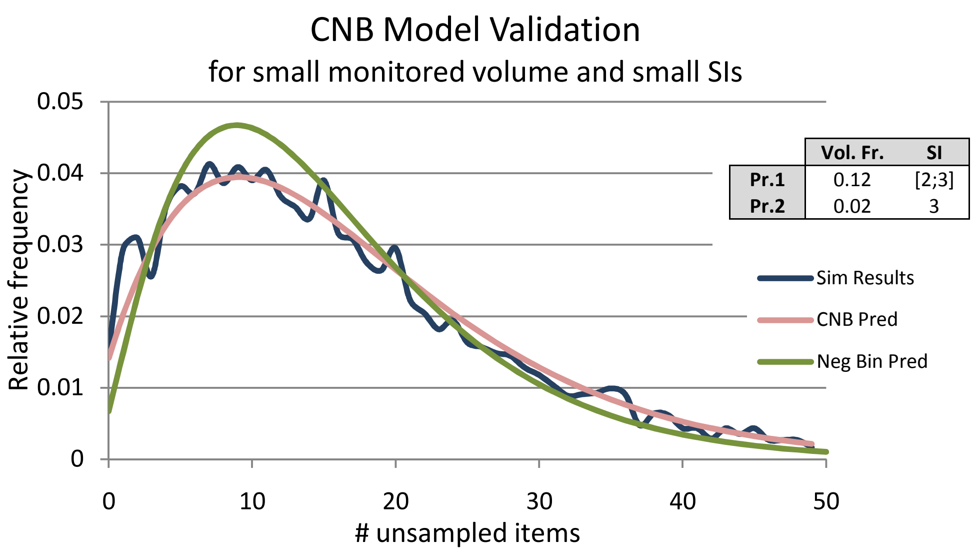


FIGURE 5.59 COMPARISON WITH THE SIMULATION RESULTS FOR SCENARIO 11.

In Figure 5.58 and Figure 5.59, the distributions predicted using the approach just described fit the simulated distribution very well. The cumulative absolute error ranges are 4.20% and 2.85% for the two scenarios, respectively. Since the only difference between these two scenarios is relative to the sampling intervals, the reduction of the error range and the maximum cumulative error for the second scenario is presumably due to the reduced sampling interval. This is reasonable for the fact that the wider the range of the number of monitored unsampled items is, the higher the associated variability in terms of time between samples is. In a wider time range, different numbers of unmonitored items can be processed and, as a consequence, the associated variability increases. Hence, the accumulation of the variability for a high sampling interval causes anomalies in the final distribution which are noticeable in the irregular pattern of the simulation result distribution in Figure 5.58. The pattern slightly smoothes out in the second scenario, as Figure 5.59 shows. The better fit for the second scenario is confirmed by the results of the runs test. For the scenario in Figure 5.58, the P-values obtained are 0.45 for the CNB prediction and 0.42 for the negative binomial prediction. The P-values for the second scenario (Figure 5.59) are 0.55 and 0.36, respectively. As is evident, for both the scenarios the shape of the distribution predicted using the CNB approach follows the shape of the distribution better than the prediction based on the negative binomial approximation.

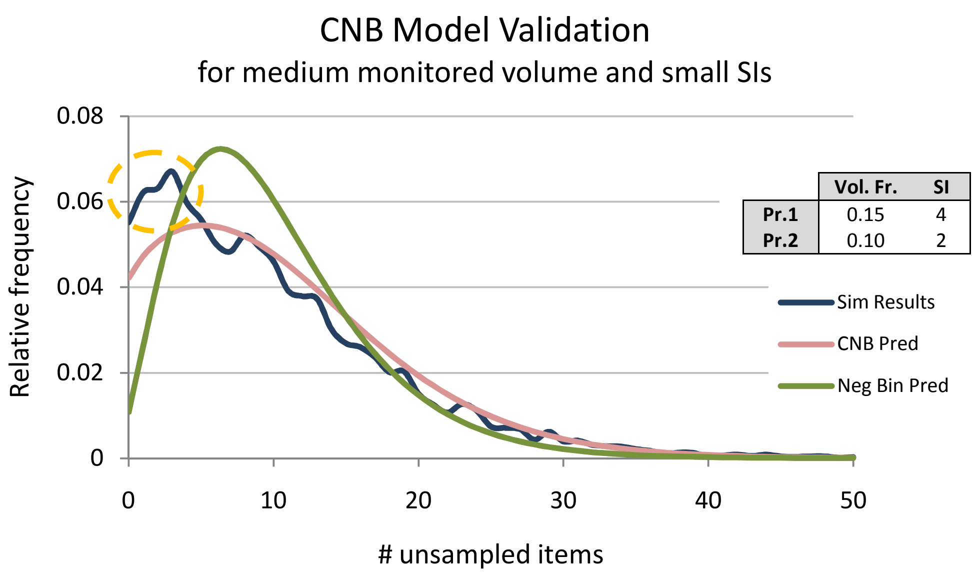


FIGURE 5.60 COMPARISON WITH THE SIMULATION RESULTS FOR SCENARIO SS5.

The goodness of fit of the predicted distribution slightly degenerates for the scenario represented in Figure 5.60. Here, the cumulative error range is 6.94%. The predicted distribution seems to underestimate the actual distribution for the lower values. In fact the peak delimited by the interval [0;5] and circled in a dashed yellow line in Figure 5.60 is not modelled by the predicted distribution which presents a smoother peak in correspondence with six unsampled items.

Figure 5.61 shows that the cumulative absolute prediction error range is affected very little by the average sampling intervals; however, the slightly higher values for the highest frequencies make the observation about the comparison between Figure 5.58 and Figure 5.59 still reasonably valid. Figure 5.61 also highlights that the monitored volume fraction impacts the quality of the prediction. Very low error ranges are obtained when the monitored products represent most of the production volume. This could be related to the greater randomness introduced by the presence of an unmonitored flow. Even though the CNB model generally provides reasonably good predictions, the poor accuracy for the low monitored volume fraction in Figure 5.61 and the failure in predicting the peak in Figure 5.60 suggest that improvements could be made to this model.

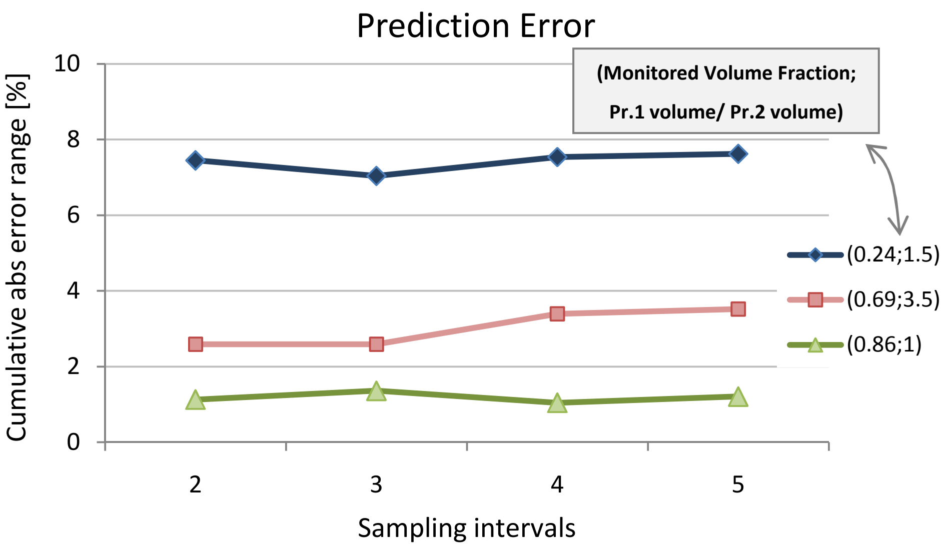


FIGURE 5.61 CUMULATIVE ABSOLUTE ERROR RANGE PATTERN WITH RESPECT TO THE SAMPLING INTERVALS (EQUAL FOR BOTH THE PRODUCTS).

5.5.3.2 *Peak correction variant*

In order to develop a correction variant to the algorithm just described, the results obtained were analysed in greater depth. The most important conclusion found was that the peak of the simulated distribution in Figure 5.60 corresponds with the domain of the associated distribution of the number of unsampled monitored items. As such, it appears that this last distribution overlaps with the distribution predicted by the algorithm and generates a slightly peakier distribution. This could be caused by the fundamental role played by the monitored products for the shape of the distribution of the number of unsampled items. Indeed, the cross flow can be considered a disturbance element for the sampling regularity; depending on the way the unmonitored flow mingles with the monitored flow, the presence of unmonitored items can either magnify the scale of the horizontal axis of the originating distribution and smooth its shape out or let the originating distribution transpire, when the number of unmonitored items processed between consecutive monitored items proves very little for the natural variability of the inter-arrival process. In order to verify if the presence of the peak in the distribution reported in Figure 5.60 is just a coincidence or a repeated pattern, different scenarios were simulated. First, different sampling intervals applied to the scenario relative to Figure 5.60 (scenario SS5) were investigated. Then, based on the observation that the prediction method provides poorer results for the cases characterised by low monitored volume fractions, different combinations of product volume fractions were also considered.

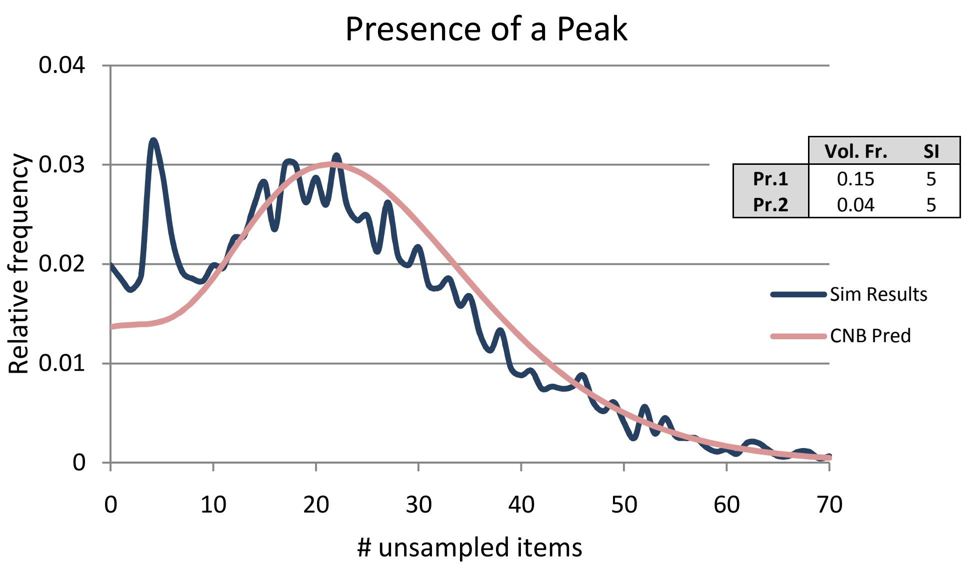


FIGURE 5.62 PEAK ANALYSIS: MISSING PREDICTION OF THE PEAK IN THE CNB MODEL (SCENARIO SS1).

The results shown in Figure 5.62 reveal the same error pattern as the one noticed in Figure 5.60. This also happened for all the other scenarios investigated. More graphs showing the systematic nature of the missing peak prediction are presented in Appendix E. In correspondence with the interval representing the domain of the associated distribution, a peak, which disrupts the smooth pattern of the distribution, arises. This suggests that a correction variant to the CNB algorithm for the peak prediction should include the distribution of the number of unsampled monitored items in the prediction algorithm. This was achieved by averaging this distribution with the distribution predicted by the CNB algorithm. As a weighting factor, it was considered that a statistic which summarises the information about the volume fractions and the sampling intervals could provide interesting results. In fact, it is likely that the impact of the distribution of the number of unsampled monitored items depends on the relevance that monitored items have in the system. As a consequence, the sampled fraction was chosen as the weighting factor, since it is related to the mentioned elements and it is immediately predictable by Equation 5.5.

The correction was applied to all the cases analysed and the results obtained show that the fit is very good and the peak is quite correctly reproduced. The cumulative absolute error ranges are 1.8%, 2.35% and 6.35%, respectively, for the results shown in Figure 5.63, Figure 5.64 and Figure 5.65.

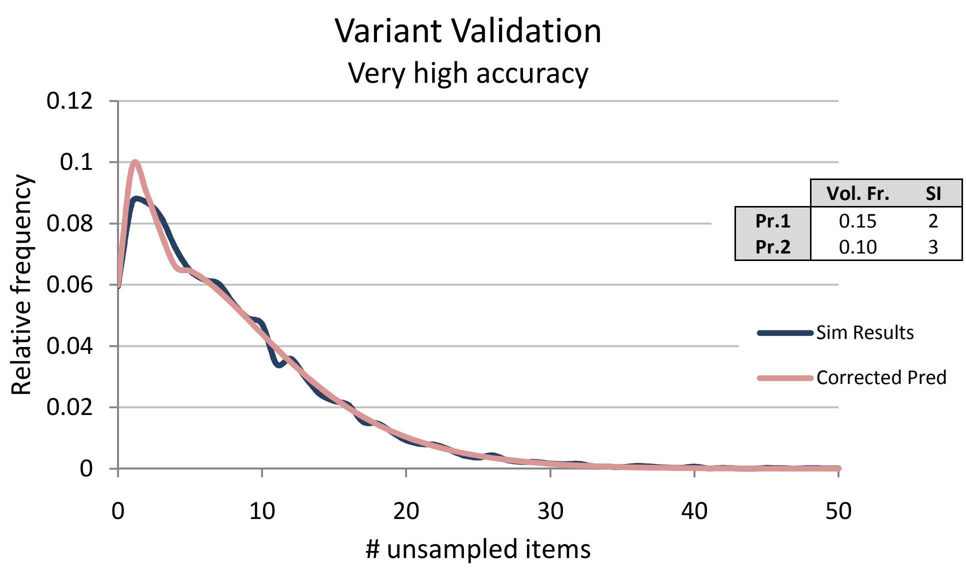


FIGURE 5.63 VARIANT VALIDATION: COMPARISON WITH SIMULATION RESULTS FOR SCENARIO SS5 (LOW SAMPLING INTERVALS).

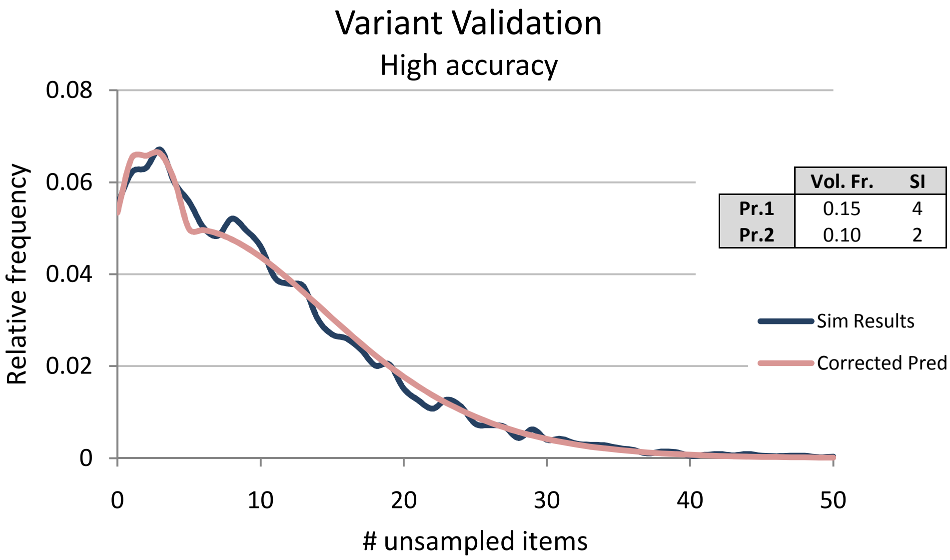


FIGURE 5.64 VARIANT VALIDATION: COMPARISON WITH SIMULATION RESULTS FOR SCENARIO SS5 (HIGH SAMPLING INTERVALS).

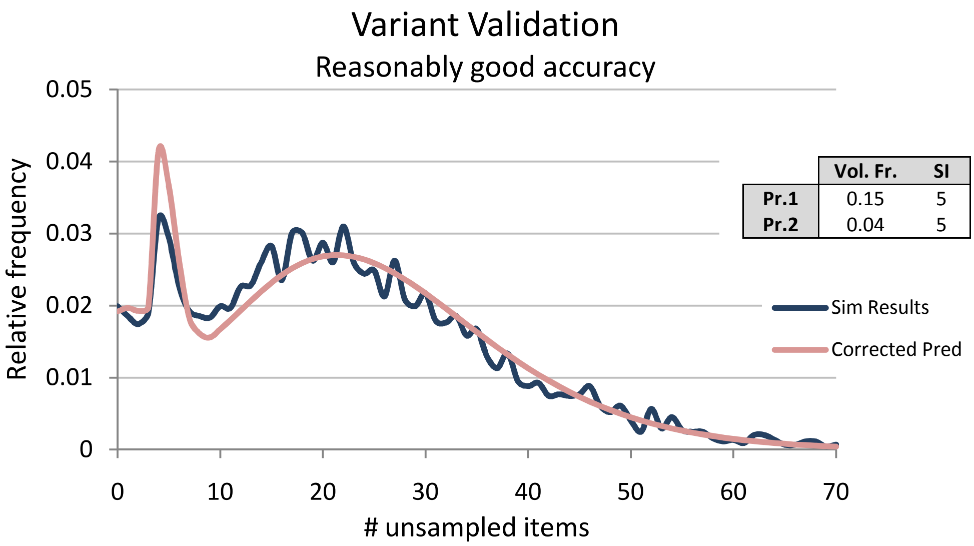


FIGURE 5.65 VARIANT VALIDATION: COMPARISON WITH SIMULATION RESULTS FOR SCENARIO SS1.

Even though the peak prediction is quite good, the results shown in Figure 5.65 suggest that a variation of the weighting factor could probably deliver beneficial effect to the prediction model. Analogous problems with the peak prediction can be noticed in the scenarios previously represented Figure 5.58 and Figure 5.59. As with the scenario in Figure 5.65, they are characterised by a very low monitored volume fraction.

The very good prediction obtained using the CNB approach suggests that any correction could worsen the goodness of fit of the predicted distribution. In fact, the results shown in Figure 5.66 and Figure 5.67 reveal that the peak predicted by the algorithm correction variant does not find an actual correspondence with the simulated distribution, at least in Figure 5.66.

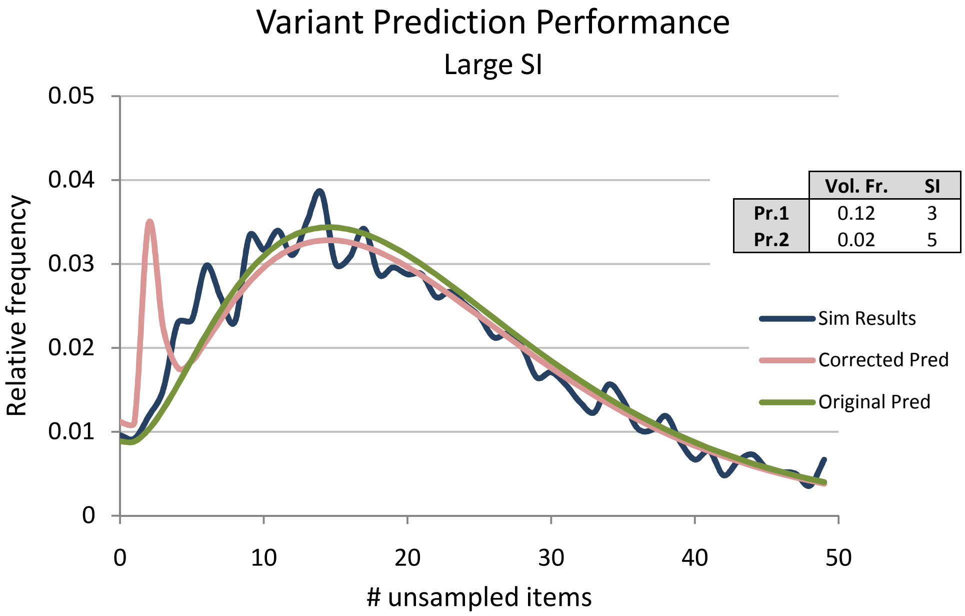


FIGURE 5.66 POOR PREDICTION RESULTS WITH THE PEAK CORRECTION VARIANT FOR LOW MONITORED FRACTIONS (SCENARIO 1).

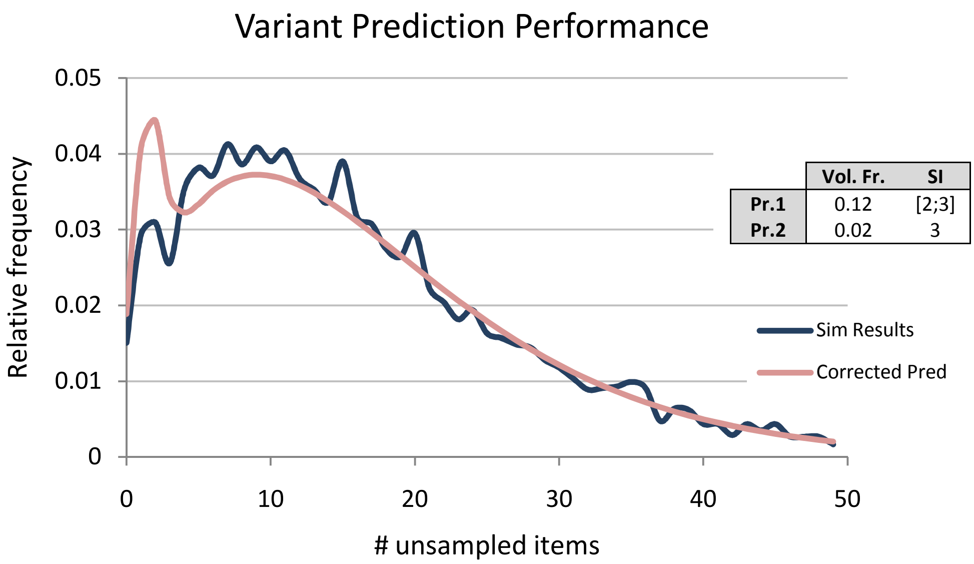


FIGURE 5.67 POOR PREDICTION RESULTS WITH THE PEAK CORRECTION VARIANT FOR LOW MONITORED FRACTIONS (SCENARIO 11).

However, in terms of prediction errors, the correction implemented still proves beneficial. In fact, the shift of the predicted distribution towards the lower values of the domain, caused by the peak presence, reduces the gap between the actual and the predicted distribution for the upper tail. The cumulative error range is reduced by 0.5% and it is equal to 3.74%. The benefits of the correction variant are confirmed by the runs tests applied to both the distributions; an increase the P-value from 0.27 to 0.45 is obtained for the correction variant which means that the corrected model follows more closely the shape of the actual distribution. As regards Figure 5.67, the actual distribution presents a peak, however, it is less prominent than the one predicted by the corrected algorithm. It looks like a lower weighting factor would be more appropriate for this case. In comparison with the CNB prediction model, the error range increases by 1% and is equal to 3.88%.

The lack of any regularity in the results obtained for the correction variant made the analysis of the error pattern a fundamental step for the understanding of the actual benefits brought by the variation. For the same scenarios considered in Figure 5.61, the cumulative absolute prediction error pattern is illustrated in Figure 5.68.

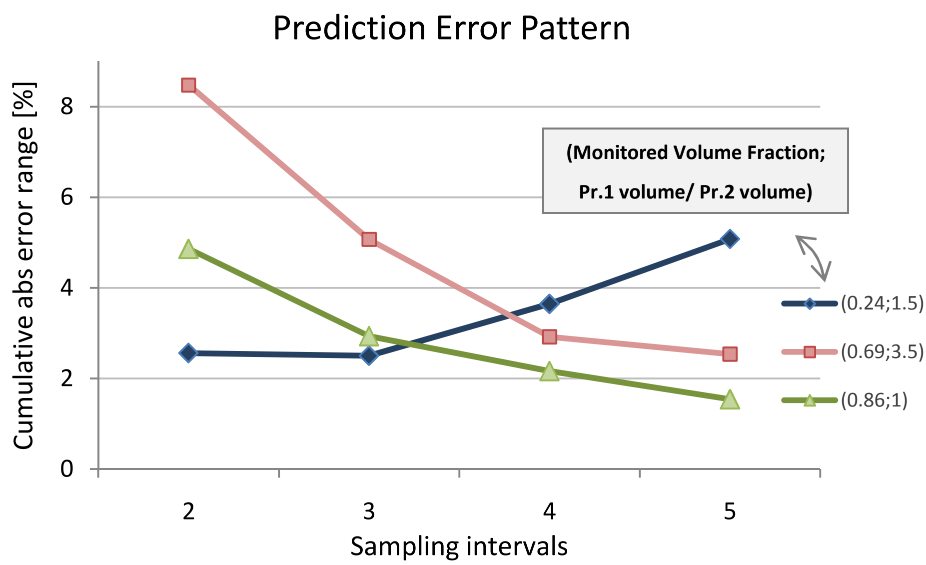


FIGURE 5.68 CUMULATIVE ABSOLUTE ERROR RANGE PATTERN FOR THE CORRECTED ALGORITHM WITH RESPECT TO THE SAMPLING INTERVALS (EQUAL FOR BOTH THE PRODUCTS).

This figure reveals that the corrected algorithm outperforms the original one only in a few cases, in terms of cumulative error ranges. The most relevant benefits are obtained for the low monitored volume fraction scenario for which the prediction errors are consistently reduced. The worst result is obtained for the lowest average sampling interval of the second scenario, for which the prediction error, increases by almost 6%. The cumulative error range, which was independent of the sampling intervals, now assumes an irregular pattern with respect to those. In fact, for the higher monitored volume fractions the error decreases with the sampling intervals; whereas for the low monitored volume fraction it increases.

Based on these considerations and those previously made on the visual comparison between the distributions, it became apparent that an optimisation search for the weighting factor could both improve the results and justify the irregular pattern of the prediction error.

The optimisation search was first based on a trial and error approach. Different weights were tested on a few scenarios and based on the best results obtained a set of weights was defined. Then, the weights in this set were applied to the other available scenarios and the best one, in terms of cumulative error range was identified. When, based on the analysis of the variation of the shape of the predicted distribution, it was noticed that weights outside the set could produce better results, the search domain was expanded.

First of all, the results obtained show that the weights which minimise the cumulative error ranges are apparently related neither to the monitored volume fraction nor to the sampling intervals. This is shown in Figure 5.69, where for each unmonitored fraction simulated the average weights over 10 different sampling intervals investigated are reported. The representation of the average value is still meaningful since the dispersion of the optimal weights relative to a particular unmonitored fraction is very limited. The pattern of the optimal weights is very irregular; however, apart from the last point, the weights are very close to each other. The anomalous value obtained for the last point is not clearly justifiable; however, it has to be said that, for this last case (scenarios 1 and 11), a simulation model slightly different from the one used in the other scenarios analysed here (Table 5.6) was considered. The model did not differ substantially, either in its design or in the shape of the input distribution used. For this reason, it is quite difficult to find any reason for the relevant variation of the weighting factor. Further investigations would be advisable to determine a prediction model for the optimal weight; there is a strong inclination to believe that the dispersion of the distribution of the number of consecutive unsampled items and the dispersion of the distribution of the number of consecutive unsampled monitored items could be related to the optimal weight. The only reason for this is based on the fact that the dispersion of those distributions is the only statistics that could eventually vary between the scenarios. In fact, unlike the mean value of the distribution, which can be easily predicted and has proved not to be related with the optimal weight, the dispersion is not directly predictable yet. It could be derived by the distribution itself, but this would mean a deadlock situation in the case where it was needed for the prediction of the distribution itself, as happens for the corrected approach.

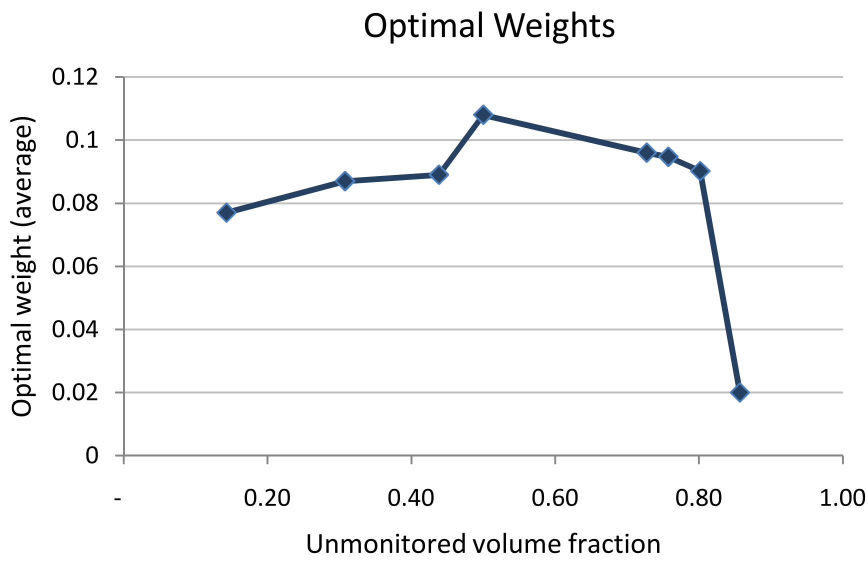


FIGURE 5.69 OPTIMAL WEIGHTS OF THE CORRECTED ALGORITHM WITH RESPECT TO THE UNMONITORED VOLUME FRACTION.

The choice to leave a deeper analysis of the optimal weight prediction as a future development of this research project is mainly based on the consideration that, even though less accurate, the original approach still proves good enough for the prediction of the distribution of the number of consecutive unsampled items.

The optimal weight significantly reduces the cumulative absolute prediction error range, such that in a few cases it is lower than 0.5%. The patterns of the cumulative error range for the scenarios also analysed in Figure 5.61 and Figure 5.68 are reported in Figure 5.70. The error is almost independent of the sampling intervals, even though a slight increasing slope is detectable. The monitored volume fraction is again a significant factor, as happened in the original approach (Figure 5.61); the scenarios with a low monitored volume fraction present a lower prediction accuracy. However, even in those cases, an average cumulative error range equal to 2.4% gives great confidence in the validity of the prediction model.

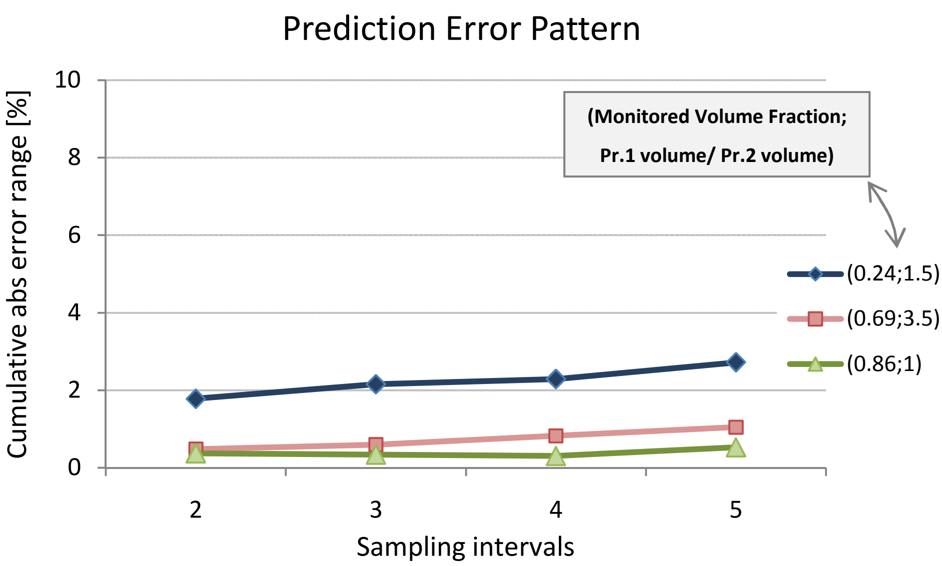


FIGURE 5.70 CUMULATIVE ABSOLUTE ERROR RANGE PATTERN WHEN OPTIMAL WEIGHTS ARE USED IN THE CNB MODEL VARIANT.

The apparent illogical error patterns in Figure 5.68 can now be justified. In fact, unlike the optimal weights which do not change with the sampling intervals, the calculated weights do. As a consequence, the prediction errors increase or reduce based on the closeness of the currently calculated weighting factor to the optimal weighting factor.

The runs test applied to the different scenarios analysed always highlighted the statistical significance of the goodness of fit for the CNB corrected predictions.

5.5.3.3 Impact of the error of the distribution in input

So far, the predicted distributions have been generated by using the simulation results to develop the distributions of the number of monitored unsampled items between consecutive samples. This choice is similar to the one followed for the previous case analysed, when only two monitored products cross the sampling station. It was suggested by the need to avoid a misleading evaluation of the prediction model as a consequence of an error accumulation effect. However, considering that for the application of the approach the distribution of the number of consecutive unsampled monitored items is needed even when no historical information is available, it is advisable to assess how the use of the predicted distribution affects the results. This was done for the scenarios in Figure 5.58 and Figure 5.59. The original algorithm, without the correction variant, was used for the prediction of the distribution of the number of unsampled items between samples.

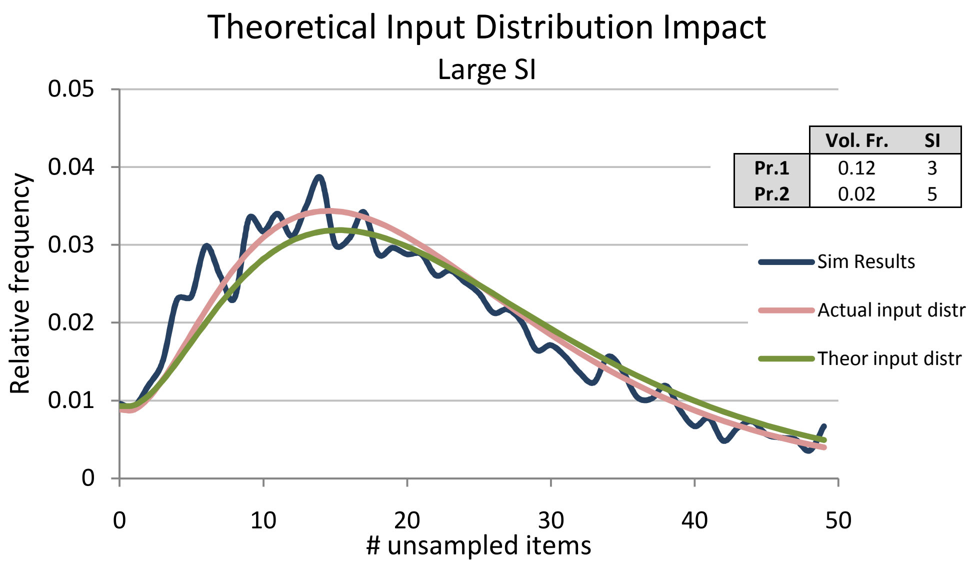


FIGURE 5.71 IMPACT OF THE PREDICTED INPUT DISTRIBUTION ON THE DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS FOR SCENARIO 1.

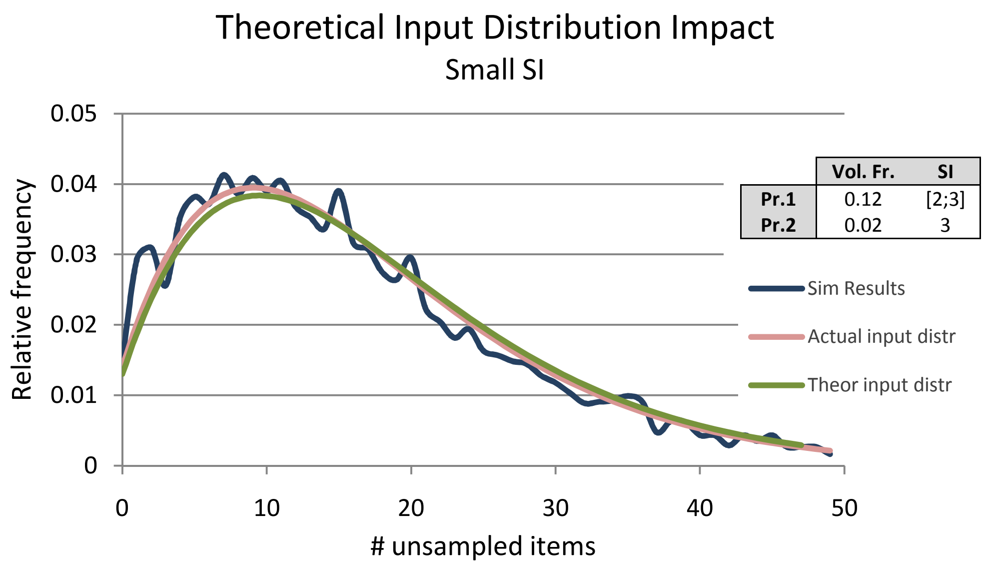


FIGURE 5.72 IMPACT OF THE PREDICTED INPUT DISTRIBUTION ON THE DISTRIBUTION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS FOR SCENARIO 11.

The agreement between the experimental and the predicted distributions still seems good enough even though a bias in the predicted distribution can be noticed (Figure 5.71 and Figure 5.72). That is due to the error pattern of the prediction model for the distribution of the number of consecutive monitored unsampled items which usually underestimate the lower tail of the actual distribution. The error ranges in the prediction of the distribution of the number of unsampled monitored items, equal to 5.23% and 3.31%, determine an increase in the prediction errors of the distribution of the number of consecutive unsampled items, whose ranges are equal to 5.56% and 4.24% for the first and second scenario, respectively. That means the error ranges are increased by 1.36% and 1.39%, respectively, which is much lower than the prediction errors of the original predicted distributions. The errors, then, are certainly transmitted but not merely summed up, which is a noticeable advantage for the validity of the approach.

5.5.4 Impact of the time related distributions

The robustness of the CNB prediction model and its correction variant was also tested with respect to variations of both shape and variability of the time related input distributions. The analysis conducted, the extra scenarios simulated and the results obtained are presented in details in Appendix F. In this section the main findings are summarised.

The importance of the hypothesis of exponentially distributed inter-arrival times for the validity of the CNB prediction model was investigated. This followed the consideration that the CNB model is based on the hypothesis of independence of consecutively processed items, which is guaranteed by the memory-less property of the exponential distribution. The results of scenarios SS2, SS7 and SS8 (Table 5.6 (p. 185)) show that when the coefficient of variation is kept unchanged (100%), lognormally distributed inter-arrival times do not impact the distribution of the number of consecutive unsampled items. The lognormal distribution was the only alternative considered to the exponential distribution since it would be very unlikely to encounter other distribution types to model inter-arrival times. It might be interesting to investigate the effect of inter-arrival time variability on the distribution of the number of unsampled items; however, at this regard, the results obtained in Section 5.4.4 supported the decision to leave this analysis for future work.

In the extra set of scenarios simulated for this analysis, the processing times were turned into deterministic times, whereas the queuing and transportation times and the availability times, that is MTBF and MTTR, were alternately modelled as exponential and lognormal distributions with different levels of variability according to the experimental plan reported in Table 5.8.

TABLE 5.8 EXPERIMENTAL DESIGN.

		Queuing & Transportation Times: Historical Data	Queuing & Transportation Times: Exp.	Queuing & Transportation Times: Lognorm. CV=50%	Queuing & Transportation Times: Lognorm. CV=150%
Availability Times	Historical Data	Sc. A	Sc.B		
	Exp.		Sc. C	Sc. D	Sc. E
	Lognorm. CV=50%		Sc. F		
	Lognorm. CV=150%		Sc. G		

The results obtained suggest that the variability and the shape of the queuing and transportation time distribution do not impact the mean and the standard deviation of the distribution of the number of consecutive unsampled items. This is probably because the distributions considered during the analysis are not significantly different from the original distributions in terms of shape. Moreover, the variability of the queuing times contributes to increase the sequence disorder effect, which, in a sampling station, is not as fundamental as in a non-sampling station for the characterisation of the distribution of the number of unsampled items.

The noticeable variation of shape for the availability times for some stations causes a significant reduction of the standard deviation of the number of consecutive unsampled items as is evident in Figure 5.73. It is worth noting that, even though the exponential distribution is commonly used for modelling MTBF and MTTR, it is not suitable when regular maintenance events are simulated. These events are characterised by an almost deterministic recurrence and the time needed to perform the corresponding operations presents a very small variability.

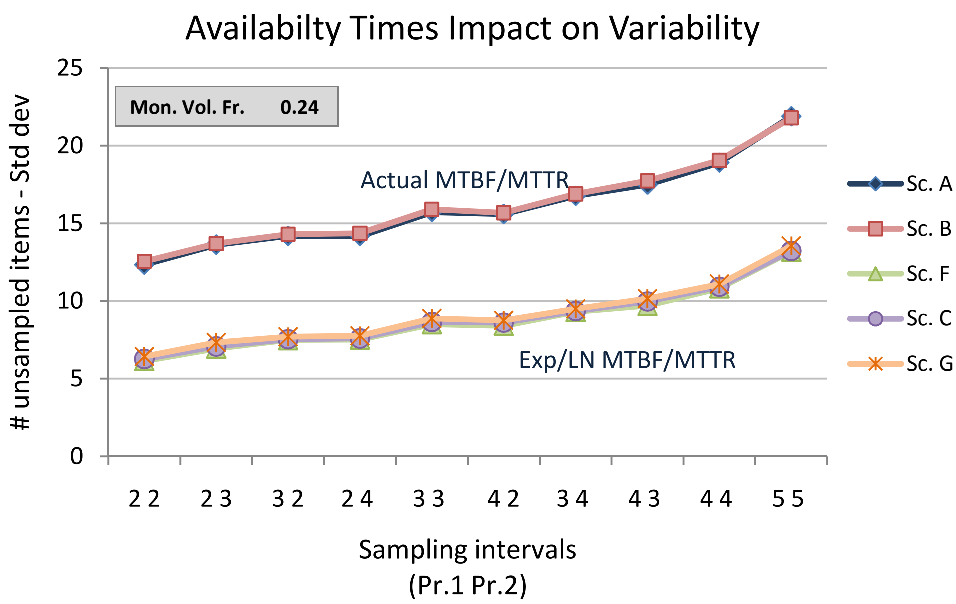


FIGURE 5.73 IMPACT OF THE SHAPE OF THE AVAILABILITY TIMES ON THE STANDARD DEVIATION OF THE NUMBER OF CONSECUTIVE UNSAMPLED ITEMS.

The variability reduction of the number of unsampled items is reflected in a little variation of the shape of the distribution of the number of consecutive unsampled items, as can be seen in Figure 5.74.

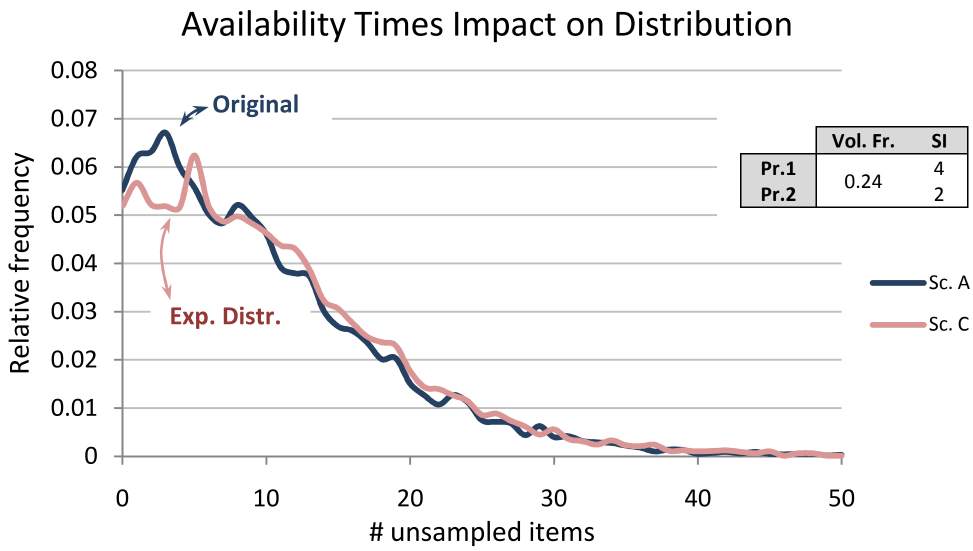


FIGURE 5.74 VARIATION OF THE SHAPE OF THE DISTRIBUTION OF THE NUMBER OF UNSAMPLED ITEMS AS THE AVAILABILITY TIMES DISTRIBUTIONS CHANGE.

The variation of both the variability and the shape of the distribution of the number of unsampled items suggest that the accuracy of the CNB prediction model could vary for the new scenarios (F, C and G) with respect to the scenarios with the actual availability times (A and B). The optimal weighting factor to be used in the correction variant would probably change as well.

Relatively to the scenario in Figure 5.75, the cumulative error range for the CNB prediction reduces from 6.94%, when actual availability times are considered, to 3.37%, when exponential availability times are used. Also the optimal weighting factor reduces from 0.08 to 0.035; the cumulative error range associated with the correction variant is as little as 1.56%.

This confirms that the peak correction variant outperforms the CNB model, provided that the optimal weight is known. The results obtained also support the opinion that the optimal weight depends on the variability of the distribution of the number of consecutive unsampled items. In fact, when the variability reduces, the accuracy of the CNB prediction model improves and, as a consequence, the weighting factor reduces. The availability of prediction formulae for the standard deviation of the number of unsampled items should ease considerations on the approximated magnitude of the weighting factor and could eventually lead to its prediction. The results obtained also suggest that the standard deviation of the number of consecutive unsampled items is affected by characteristics of the time related input distributions which are ignored for the prediction of both the mean value and the distribution of the number of consecutive unsampled items.

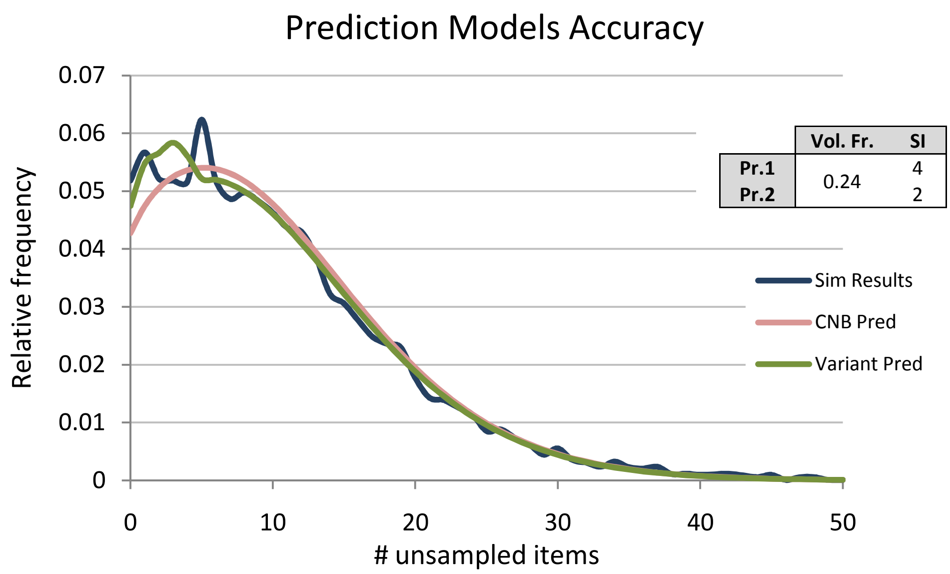


FIGURE 5.75 PREDICTION EFFECTIVENESS OF THE CNB MODEL AND ITS VARIANT FOR EXPONENTIALLY DISTRIBUTED AVAILABILITY TIMES (SCENARIO C).

5.6 Applications in industry

When it was initiated, one of the main motivations for this research project was to develop a model for both predicting the risk associated with a sampling strategy and setting the sampling strategy parameters which would guarantee operating the strategy with a quality risk level lower than a predetermined level.

The development of the prediction models of the distribution of the number of unsampled items between consecutive samples addresses these issues. In fact, if the number of consecutive unsampled items is chosen as a quality risk related performance measure, the possibility to predict its distribution in different operating conditions allows any risk level to be quantified in terms of maximum number of consecutive unsampled items exposed to that level of risk. At the same time, the distributions should make it possible to derive the sampling strategy parameters once the maximum acceptable number of items exposed to a certain risk of not been sampled is set.

Given that the problem of predicting the risk associated with a sampling strategy and the problem of setting the sampling strategy parameters able to guarantee a given quality risk level are dual, it is reasonable that the approaches followed to solve them are approximately based on symmetric procedures. In the next sections the problems will be separately considered and a few applications will be illustrated.

5.6.1 Quality risk associated with a sampling strategy

The first problem considered regards the definition of a quality risk level associated with a sampling strategy based on deterministic sampling intervals in one step of a linear production segment. The prediction models developed in this chapter can produce a straightforward answer when the number of consecutive items exposed to the risk of not being sampled is considered as an effective measure of this risk. The confidence that the number of consecutive unsampled items could be extensively used as an appropriate measure is based on the consideration that, its complementary measure, which is the risk exposure time, does not always properly quantify the magnitude of the production losses as a consequence of a quality failure. In fact, it can happen that, due to low machine utilisation, a very high time from the introduction of the quality failure to its detection corresponds with a very low number of defective items being produced. Hence, the information yielded by the time measure could prove misleading from a quality risk viewpoint.

The quantification of the quality risk in terms of maximum number of unsampled items between consecutive samples in a production station is simply based on the calculation of the percentile of the distribution of the number of consecutive unsampled items corresponding with a cumulative probability equal to the risk level that the quality management intends to take as a reference level. The distribution to be considered obviously depends on the station type, whether a non sampling or a sampling station, and, in the last case, on the combination of product flows that cross the station.

For example, if management agrees to assume as a quality risk measure the risk of not consecutively sampling items at a 90% confidence level, the 90th percentile of the distribution of the number of consecutive unsampled items is the solution to the problem. That means that only in 10% of cases, the maximum number of consecutive unsampled items will be greater than the evaluated quality risk measure.

In order to show how to apply this procedure in the different stations of a production segment, the following case will be considered. Assume that the segment is crossed by two monitored products A and B, which have exponentially distributed inter-arrival times, with mean equal to 2 and 4 [time unit/items], respectively; all the stations also receive cross flow items with an average inter-arrival time equal to 0.5 time units. The sampling intervals are set to 4 items and 2 items for the two products. In this case, the monitored volume fraction is equal to 27% and the sampled fraction equal to 9%. The average sampling interval of the global monitored flow is 3 items. For the non-sampling station the predicted distribution of the number of consecutive unsampled items is geometric with parameter p equal to the sampled fraction.

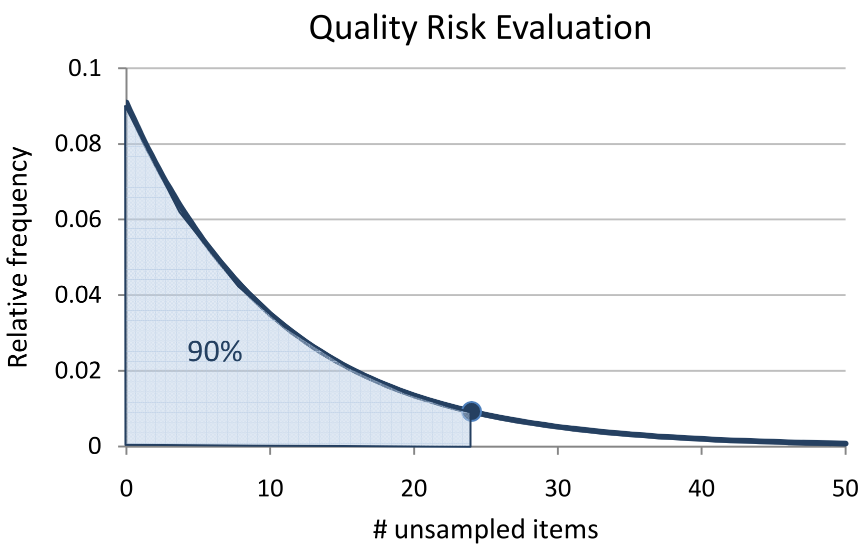


FIGURE 5.76 QUALITY RISK IN A NON-SAMPLING STATION IN TERMS OF NUMBER OF CONSECUTIVE UNSAMPLED ITEMS AT A 90% CONFIDENCE LEVEL.

Figure 5.76 shows the region of the predicted distribution of the number of consecutive unsampled items in a non-sampling station which corresponds with a 90% cumulative probability. The value on the x axis delimiting the region consists of the 90th percentile of the distribution; it can also be interpreted as the quantification of the 90% quality risk in terms of number of items between samples. In the example considered, with a sampled fraction of 9%, there exists a 90% risk of not consecutively sampling 24 items at most. This also means that there is only a 10% chance of having more than 24 items consecutively unsampled.

In the case of a sampling station, the prediction based on the original approach, without the peak correction, suggests that the risk corresponding with a 90% confidence level is equal to 20 consecutive unsampled items (Figure 5.77, pink line). The risk reduction is expectable since the distribution of the sampling station is characterised by a lower variability than the non-sampling station one, whereas the mean value is the same for all the stations. This results in a peakier distribution for the sampling station.

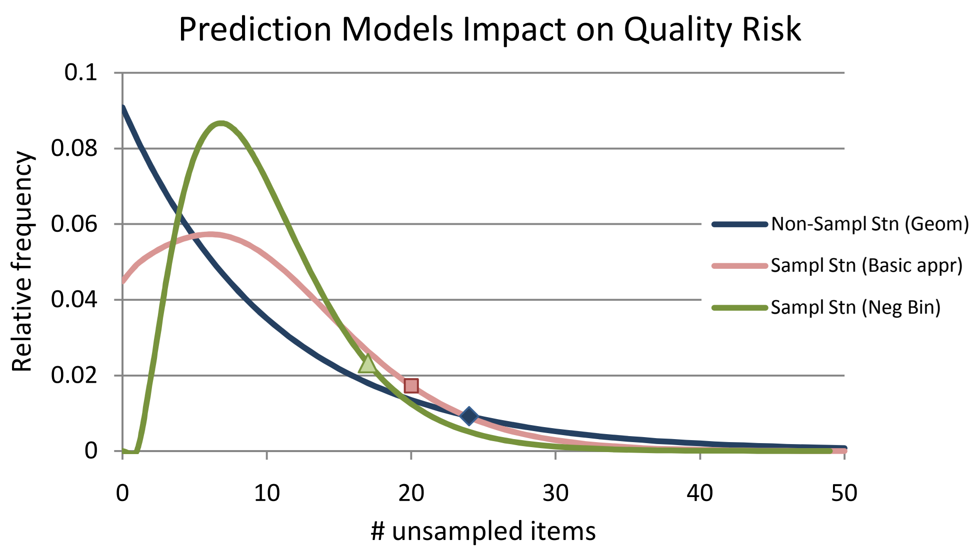


FIGURE 5.77 DISTRIBUTIONS OF THE NUMBER OF UNSAMPLED ITEMS FOR THE PREDICTION OF THE QUALITY RISK.

The negative binomial approximation for the distribution of the number of unsampled items in a sampling-station crossed by an unmonitored flow determines a further reduction of the risk. In fact, the number of consecutive items exposed to the risk of not being sampled is 17 at a 90% confidence level (Figure 5.77, green line). A visual comparison between the three predicted distributions used in this analysis is proposed in Figure 5.77.

5.6.2 Sampling strategy with quality risk constraints

The other problem to be solved consists of the definition of the sampling parameters which satisfy the imposed quality risk constraints. For example, if management wishes to keep the maximum number of items exposed to the risk of not being sampled lower than a given value, at a certain confidence level, the prediction model should be able to suggest the smallest sampled fraction needed to meet the quality risk specifications. Based on that, the monitored product fractions and their respective sampling intervals can be derived.

The simplest way to solve this problem is to consider the geometric prediction model for the distribution of the number of consecutive unsampled items as a reference, independently of the nature of the station, whether a non-sampling or a sampling one. The reason for this is twofold. First, unlike the other prediction models developed, the geometric model is characterised by only one parameter, which is the sampled fraction; this means that the minimum possible number of unknown variables is involved in the procedure. Second, as the example in the previous paragraph showed, the quality risk measures estimated by the geometric distribution prove more conservative than the ones generated by the other prediction approaches, in the sense that, due to the higher variability of the geometric distribution in comparison with the negative binomial distribution, the percentiles associated with it are always greater than those generated by the other prediction approaches. This means that if the required sampled fraction is calculated based on the geometric prediction, it would be the minimum sampled fraction which guarantees the constraint validity in any station of the segment.

As regards the higher variability of the distribution of the number of unsampled items in a non-sampling station in comparison with a sampling station, there are different ways to prove it. First, it is immediately observable that if there is a deterministic element in the sampling station, the variability of the resulting distribution should be less than the variability of a station where the sampling strategy is apparently random. Second, a confirmation of this intuition was obtained with the simulation results. Third, also as a consequence of this, the prediction models reflect this difference in variability. Assuming that the negative binomial distribution is adopted to approximate the distribution of the number of unsampled items in a sampling station, the difference in the variability can be mathematically proved. In fact, given that the sampled fraction, p_G, which is the parameter of the geometric model, and the monitored fraction, p_{NB}, which is one of the two parameters of the negative binomial model, are related to each other by means of the average sampling interval, s, which is the second parameter of the negative binomial model, then:

$$p_G = \frac{p_{NB}}{s} \tag{5.18}$$

Consequently, the variances of the two distributions can be expressed in terms of the same proportion. In particular, the variance of the geometric distribution can be expressed as follows:

$$\begin{aligned}\sigma_G^2 &= \frac{1-p_G}{p_G^2} \\ &= \frac{1-\frac{p_{NB}}{s}}{\frac{p_{NB}^2}{s^2}} \\ &= s\,\frac{s-p_{NB}}{p_{NB}^2} \\ &= \frac{s-p_{NB}}{1-p_{NB}}\,\sigma_{NB}^2\end{aligned} \tag{5.19}$$

Since $s \geq 1$, the equivalence in Equation 5.19 proves that the geometric model has a higher variability than any negative binomial distribution with the same mean as the geometric distribution. Equation 5.19 also reveals that, keeping the same mean, the variability of the negative binomial distribution increases with the average sampling interval. This observation is useful when the sampling strategy parameters have to be set.

Once defined the quality constraint, the minimum necessary sampled fraction with respect to the constraints can be found using an iterative approach. Based on this, the sampling intervals and, whenever the production plans flexibility allows it, the volume fractions for the monitored products can be set so that the calculated sampled fraction is obtained. This can be formally expressed as follows:

$$\sum_i \frac{p_i}{s_i} = p_G \qquad \textbf{(5.20)}$$

where the index i refers to all the products chosen to be monitored and p_i and s_i are the volume fraction and the sampling interval of the i^{th} monitored product type, respectively. Since there is only one condition to be satisfied, only one unknown parameter, among the p_i's and s_i's can be determined; the rest of them can be set at will, that is, Equation 5.20 presents $2n$-1 degrees of freedom, if n is the number of monitored products. It is worth noting that the product mix is usually set based on the product demand, so the volume fractions, p_i, are most likely already known when the sampling strategy is to be determined. This reduces the degrees of freedom in Equation 5.20 to n-1.

The observation previously made about the characteristics of the negative binomial distribution variability is useful here to make decisions about the sampling strategy parameters. In order to avoid a great variability of the number of the consecutive unsampled items in the sampling station, whenever possible, it is more convenient to increase the monitored volume rather than operate with a large volume of unmonitored flow. Even though with a greater monitored volume the average sampling intervals increase, this does not generally cause an increase in the variability as Equation 5.19 could suggest; in fact, the effect of the higher monitored proportions proves more beneficial. A greater variability the higher percentiles of the distribution shifts towards greater values; this means, the quality risk, even though lower than the risk in the non-sampling stations, is higher than it could be.

The derived sampling plan should be tested again using the appropriate prediction model for the distribution of the number of unsampled items in a sampling station to ensure that the quality constraints are respected.

For example, if management is looking for a sampling strategy so that the maximum number of items exposed to the risk of not being consecutively sampled is 10, at a 90% confidence level, a search can be done on the sampled fraction first. Starting from a reasonably low sampled fraction, for instance 0.1, based on the results obtained by applying the inverse of the geometric distribution function, the proposed sampled fraction can be increased or decreased. The increment can be decided based on the desired accuracy. In this case, after a few iterations, it was found that the minimum sampled fraction which satisfies the quality constraint is 0.19. Any sampled fraction equal to or greater than 0.19 should guarantee the conditions imposed by management. Supposing that, due to available monitoring capacity, the management decides to operate with a 20% sampled fraction and 4 products flowed through a station, different sampling parameter combinations would be possible. In the case where each product is produced in the same volume, that is 25% volume fraction for each product, using the equivalence in Equation 5.21, the decision to sample the 4 products with a sampling interval equal to 5 items should have similar effects as a decision of sampling 3 products with sampling intervals equal to 3.75 items, or 2 products with a sampling interval equal to 2.5 items, or only one product every 1.25 items. This equivalence is in terms of average number of unsampled items. The non-integer sampling intervals can be obtained using ordered sequences of integer sampling intervals (See section 5.5.2.3). The only differences are related to the number of consecutive unsampled items distribution variability, which should increase with the unmonitored volume fraction. This means, when the monitored fraction is 75%, the associated 90^{th} percentile will be less than the 90^{th} percentile associated with the option with a monitored fraction equal to 25%. However, for any option considered, the risk measure in the sampling station should prove less than 10 items not consecutively sampled.

This was proven using the negative binomial approximation for the distribution of the number of unsampled item in a sampling station. For a 75% monitored volume, the quality risk measure at a 90% confidence level is 5.75 items; for a 25% monitored fraction the same measure is 9.25 items, which is very close to the non-sampling station measure (Figure 5.78).

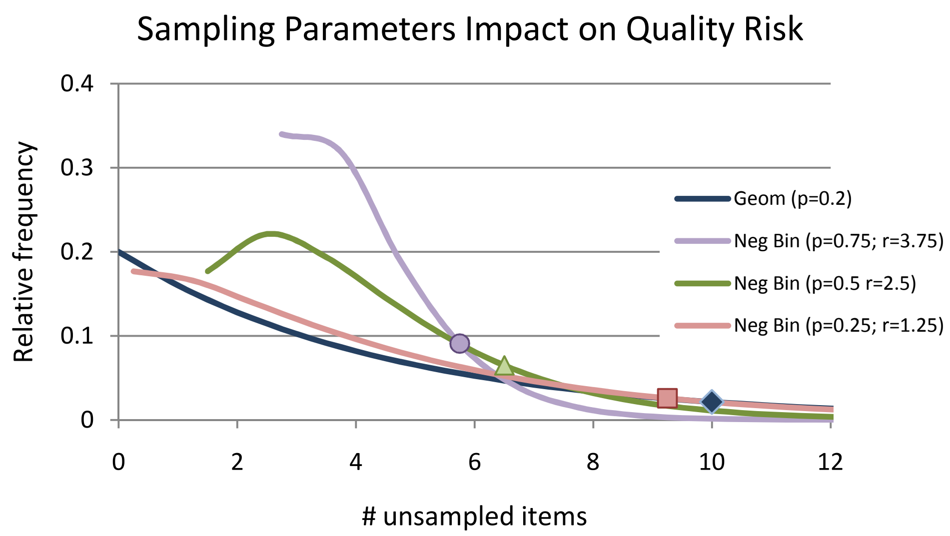


FIGURE 5.78 IMPACT OF THE SAMPLING STRATEGY PARAMETERS ON THE NUMBER OF MAXIMUM ITEMS AT RISK (90% CONFIDENCE).

5.6.3 Industrial applicability

The models developed in this research and the approaches suggested for deriving both quality risk estimations and optimal sampling parameters have been illustrated to the quality personnel of the company supporting this work. They deem the prediction models of the distribution of the number of unsampled items (Sections 5.4 and 5.5) and the considerations on the model applications (Section 5.6) very interesting and fully responsive to their initial questions. Based on their understanding of the sampling process and on analyses of historical data, the prediction models are considered capable of capturing the actual patterns of the number of items between samples in both sampling and non-sampling stations and provide realistic estimates of the quality risk.

In particular, the simple structure of the geometric model and its versatility to provide conservative quality risk estimations for both any station and any product flow combination positively impressed the management who are strongly considering the idea of using this model to support quality control related decisions. The typical inertia of industrial companies in adopting approaches consolidated by tradition is making the implementation process of the models developed in this research quite slow. In the opinion of the quality staff, the practicality of the geometric model and the limited information required in input represent key elements for attracting the interest of the industrial world and facilitate a successful implementation of this model.

The results obtained and the models developed here can be applied to a wide industrial domain. In this research, the simulation model used for the development and validation of the different prediction models has been based on a real production segment of a particular company. Real data from that company have been initially used to study and simulate the system behaviour. However, this does not necessarily mean that the structure of the simulation model and the results obtained are peculiar of the company supporting this research and applicable exclusively to it. The simulation model developed can be considered a generic model of production systems with a serial structure and parallel machines operating in consecutive stations. This type of systems is very common in industry and quite complex to analyse [11].

The production system layout is not the only element needed for the applicability of the models developed. The structure of the sampling strategy also plays a fundamental role for the application of the results found. For a serial-parallel multi-stage system, whenever the different products (or even processes) which flow through the system are sampled on a regular and deterministic basis, the prediction models developed can be considered valid. It is believed that the regularity of the sampling plan is fundamental for the validity of the prediction models for the distribution of the number of consecutive unsampled items at the machine of a sampling station. For the negative binomial model a deterministic sampling interval is essential to guarantee a non-approximated application of the model itself; this was discussed in Section 5.5.2, where the use of the average sampling interval in place of the actual ones was investigated. As regard the enumeration approach, this was developed and tested for deterministic sampling intervals; its generalisation to repetitive sequences of deterministic sampling intervals was illustrated in Section 5.5.2.3. However, due to its combinatorial nature, extending the application of the enumeration approach to random sequences of sampling intervals could prove not efficient or even not feasible. Provided that the distribution of the sampling intervals is known or derivable, the CNB model is applicable to scenarios characterised by random sampling. This statement is based on the consideration that the distribution of the monitored unsampled items, which is compound with the negative binomial distribution, is at all effects the distribution of the global sampling intervals. If this distribution was available *a priori,* the CNB model would be applicable independent of the number of monitored products in the system.

The nature of the prediction model developed for the non-sampling stations supports the belief that this model can be applied to more general scenarios than the ones considered in this thesis. Being the distribution parameter exclusively based on the sampled fraction, the validity of the geometric model could prove robust to variations of the sampling strategy if the system is characterised by an adequate level of randomness. This is to guarantee that the sequence disorder effect, which has proved fundamental for the validity of the model (Section 5.4.4), affects the system at such a level that any particular sampling pattern in the sampling station is turned into a memory-less random pattern in a non-sampling station.

The structure of the segment modelled and the location of the sampling station as the last station of the production segment suggest that the models developed have the potential to be applied in industrial environments where final inspection is performed. In order to reduce external failure costs, most companies perform a final quality check so that the number of defective products delivered to customers can be minimised. Final inspections are usually implemented prior to packaging in the pharmaceutical industry [149] as well as in the electronic [17, 150] industry. The automotive industry [151] and other large consumption goods industries [64, 117-119] also implement final inspection. In particular, the concept of final inspection can be generalised to inspections performed at the end of each cycle in re-entrant production systems such as semiconductor fabrication facilities [66, 152].

On the other hand, the applicability of the models is not confined to systems for which the inspection is located at the end of the production cycle. In Section 5.4.4, when the first station in the segment was chosen as a sampling station, similar results were obtained in comparison with the case of the sampling station located at the end of the segment. As a consequence, the observations above on the validity of the models still apply to multi-stage systems that adopt inspection location strategies alternative or complementary to the implementation of a final inspection.

Finally, the models developed can be applied to both batch and flow production. The systems analysed focused on the production of single items; however, if batches are considered in place of items nothing would change. In this case, the sampling intervals would refer to batches rather than items and the distributions will describe the number of consecutive unsampled batches rather than items. The modality with which inspection is performed within a batch, whether a screening or a random sampling, is not relevant for this analysis since it will not impact the validity of the models. This is because, for the objective of this research, the inspection within a batch will only support the evaluation of the quality level of the production, that is, it will be used to assess if the batch is defective or not. In other words, any further quality information that the inspection within the batch can provide, such as item-to-item variability or the presence of defect patterns within a batch (item), is certainly valuable from a quality perspective but it is irrelevant for the purpose of this research. Indeed, this information is not useful for the estimation of the risk of not sampling a long sequence of batches at any machine of the production system, so that the quality status of both the machines and the production process can be inferred and eventual quality failures can be detected in a short period of time.

5.7 Conclusions

In this chapter the quality risk related performance measures investigated in the previous chapter by means of a simulation approach, were analytically analysed. The parameters with the most influence were first individuated and, based on them, a formula for the prediction of the average values of both the measures was derived.

Although formally interesting for giving a mathematical shape to the relationship between quality control/production system design and quality risk, the formulae for the average values do not support a quantification of the quality risk. When assessing the efficacy of a sampling strategy, information on its worst performances is more significant than information on its average performances. In order to infer the maximum exposure to the risk of not continuously monitoring a machine in the segment, the analysis of the distributions of the quality risk related performance measures proves fundamental. The number of unsampled items between consecutive samples was further analysed to derive stochastic prediction models. A preliminary study showed that, keeping the same mean, this distribution changes shape according to the nature of the station. In particular, when a non-sampling station is considered, the distribution tends to assume an exponential pattern, whereas, for a sampling station, it has a skewed bell shape.

The non-sampling station case was considered first. The memory-less characteristics of the item departure process from any of the machines of a non-sampling station suggests that the sequence of the products being consecutively processed is random. As a consequence, the event of processing an item which will be chosen as a sample can be considered a Bernoulli trial with probability of success equal to the sampled fraction. This means that the distribution of the number of unsampled items between consecutive samples follows a geometric distribution with a parameter easily predicted by the formula for the average number of consecutive unsampled items. This prediction model is valid for any product flow combination, since it is only based on the difference between sampled and unsampled items and their relative proportions.

The validity of the geometric prediction model was also tested against the simulation results obtained using a basic production segment simulation model. That allowed a better understanding of the relevance of the multiple stream and the sequence disorder effects for the applicability of the geometric model.

As regards the sampling station, the analysis did not prove very straightforward. However, it was evident that conducting separate analyses according to the product flow combinations would help to make the investigation easier. Product flow combinations involving up to two monitored products with or without cross flow were considered and the prediction models for each of them were derived. In addition, even though it has not been tried yet, it is very reasonable to assume that the approach defined for two monitored products can be extended to situations involving three or more monitored products.

The simplest case to investigate was the case characterised by the presence of one monitored product and an unmonitored product flow. In its structure, it resembled the non-sampling station case. In fact, the sequence of processed items whether being monitored items or not, can still be considered random. Hence, the distribution of the number of unsampled items still belongs to the family of distributions based on the concept of Bernoulli trials. In this particular case, the negative binomial distribution offered the solution to the problem.

When two monitored products are considered, the solution becomes more complicated and an enumerative approach is the most direct way to derive a distribution prediction. This approach proved computational expensive and further investigation would be advisable to determine the nature of the systematic error which occurs when no unmonitored flow crosses the sampling station. Nonetheless, the predictions, compared with the simulation results, are reasonably accurate.

The introduction of the unmonitored flow was analysed by combining the prediction models developed for the previous cases. A model mainly based on averaging negative binomial distributions (CNB model) was derived and a variant was also proposed. In general, this prediction proved very accurate and quite robust to eventual errors introduced with some input parameters. The variant usually outperforms the CNB approach; however, a preliminary search is required to determine the optimal weighting factor to be used.

The impact of the shape and variability of the distribution of the time related input parameters on the distribution of the number of consecutive unsampled items in a sampling station was also investigated. It emerged that the mean of the distribution is not affected by any change, confirming the robustness of the prediction formula for the average number of consecutive unsampled items. On the contrary, the standard deviation is more vulnerable to the major changes in the system dynamics, such as the way of modelling shut down events. This has an impact on the shape of the output distributions, which appear less peaky as the standard deviation decreases.

The results obtained in this last analysis also suggested that the weighting factor used in the correction variant to the CNB prediction model might be related to the variability of the number of consecutive unsampled items, which at the moment is not yet predictable. This supposition is based on the fact that the only element changing when the optimum weighting factor varies consists of the standard deviation of the distribution of the number of consecutive unsampled items.

Finally, it was shown how to use the prediction models developed to quantify the quality risk associated with a sampling strategy for a given confidence level. The prediction models can also support the choice of sampling strategy parameters able to guarantee an actual quality risk lower than a predetermined threshold value. It was shown that the geometric model, apart from being the easiest model to use, provides conservative sampling parameters so that the quality risk threshold is respected in all the stations of the production line.

Chapter VI

Discussion

6.1 Results summary and discussion

In an industrial environment characterised by high production costs and demand for high quality, reducing the risk of producing low quality products is fundamental. The capability of predicting some quality risk measurements can be very helpful in assisting the decision process when the sampling parameters have to be set. In particular, in a complex manufacturing environment, such as a flexible manufacturing system with stations provided with machines operating in parallel, even in the presence of a deterministic sampling policy for predefined products in one of the stations, the risk assessment is not trivial due to different complexity factors. The combination of a serial flow through a set of stations and a cross flow in each station, different sampling frequencies for different products, random routing policies, and the randomness of the cycle time represent some elements which complicate the quality risk analysis. All these factors contribute to turn a deterministic sampling plan into a random sampling plan in all the stations, including the sampling stations when analysed from a global flow perspective.

For the non-sampling stations, the complexity factors can be summarised into two fundamental effects, recently introduced in the manufacturing research fields [17] and which have not been investigated in great depth so far by the research community. They are known as the sequence disorder and the multiple stream effects. The latter is a mere consequence of the presence of parallel machines in the operating stations and the randomness of routing policies; the multiple stream effect also contributes to the sequence disorder effect, which is mainly caused by the randomness of the cycle time. The absence of a logical relationship between the machines in the different stations, for the multiple stream effect, and the variation of the sequence order with which items move out from the machines at the different production steps, represent the reasons why the sampling strategy investigated in this research loses its deterministic property even in a single product situation.

As regards the sampling stations, the sequence disorder and the multiple stream effects have very little to do with the randomness that characterises the sampling strategy when the global flow is considered. In this case, the merging of different independent product flows and the randomness of their arrival process at the machines of the station is responsible for the loss of the deterministic property of the sampling strategy.

Independently of its origin, this randomness is the main cause of difficulty in controlling and predicting the quality risk associated with a particular sampling strategy. For this research two measures of effectiveness of sampling were investigated. They consider the risk of not continuously monitoring a particular machine in a production segment from the perspectives of the time and the number of processed items.

In spite of the increasing attention paid by many researchers to quality issues, to the author's knowledge, very few papers in the literature deal with the quality risk assessment of a sampling strategy in terms of the number of items between samples and the time between samples. Even though an analogy exists between the measures considered in this thesis and the time between events, which has been more extensively investigated, the analysis of the time between events is limited to the analysis of the control chart associated with it [102, 103]. There is no investigation concerning the monitoring efficacy of the time between events in the stations not directly monitored by the chart. Similarly, the analyses involving the multiple stream and the sequence disorder effects focus on the enhancement of the quality control chart performances in terms of reduction of false alarms, particularly in the stations upstream of the sampling station [17, 34]. As a consequence of the lack of very specific references, the literature review only tries to introduce in a very generic way the research fields involved in this project. Moreover, only a few papers are cited throughout the remainder of the thesis, since it was difficult to find appropriate references which could support or confute the results found, even if only in analogy.

There are three primary contributions of this research. First, the impact of some control parameters on the distributions of both the measures was investigated. Second, a few prediction models for the distribution of the number of unsampled items between consecutive samples under different operating conditions were developed. Finally, a possible way to assess the quality risk associated with a sampling strategy in terms of number of consecutive unsampled items was suggested along with a procedure to determine the sampling parameters which guarantee to operate with an acceptable quality risk level. These three contributions are progressively related with each other. The first analysis, by individuating the parameters that mostly affected the performance measures, narrowed the domain of the parameters eligible for being included in the prediction models built in the second analysis. Based on these models, the procedure for the quality risk assessment was developed.

A simulation approach was used in support to the first part of the research (Chapter 3). The number of parallel machines in a station, the WIP-turn and the sampling intervals of two products were analysed in terms of their impact on both the performance measures considered (Chapter 4). The rationale behind the choice of these parameters was the intention to cover the analysis from the perspective of the system configuration, the line speed and the sampling parameters. The consideration of production system design issues, such as the line speed and the line configuration, during the analysis of quality related issues proves particularly interesting from a research viewpoint. It follows the suggestion by Inman et al. [21] which highlighted the need for considering the mutual impact that production system design and quality issues have on each other. In this study, Inman's proposal is reinterpreted and its horizons are widened. The intersection between productivity and quality is analysed here from the perspective of the quality control rather than from the one of the quality level of the items produced; moreover, the concept of line speed includes queuing & transportation times rather than processing times only as in Inman et al . [21].

The results obtained reveal that line speed increase does not necessarily correspond with a quicker monitoring capability of the sampling strategy. In general, the time between samples is affected by parameter changes which impact the inter-arrival time at a machine level. Independent of the strategy with which the increase of the inter-arrival time at a machine level is obtained, by either increasing the inter-arrival time at a station level or increasing the number of machine in a station, the time between samples proportionally increases. The results also suggest that the most direct way to reduce the number of items between samples and the time between samples is to reduce the sampling intervals. An analogous effect can be obtained by increasing the monitored volume fraction.

The monitoring performances of the sampling strategy analysed here were also assessed in terms of detection responsiveness to quality failures. The introduction of defects following a quality failure was simulated in different fashions; different maintenance scenarios were also considered. The analysis highlighted that an intermittent production of defects delays the detection of the quality problem. The absolute worst case scenario was registered when, as a consequence of routine maintenance operations, partial corrective actions were performed on a machine which was experiencing an undetected failure. The partial restoration of the machine only causes a reduction of the production frequency of defective items that, as suggested before, is very deleterious from a quality failure detection point of view.

The second part of the analysis focused on the development of analytical models which could predict the performance measures derived by means of simulation (Chapter 5). In this part of the analysis, a hybrid approach was considered [139]. New simulation models were also built in order to investigate the validity and the robustness of the proposed prediction models.

Based on the results available from the previous analyses, the formulae for the average value of both the performance measures were derived. The time between samples proved to be proportional to the number of parallel machines in a station, which impacts the inter-arrival time at a machine level, and inversely proportional to the number of samples per unit time. This last parameter can be expressed in terms of the inter-arrival time at a station level of the monitored product types and their corresponding sampling intervals. It is immediate to derive that a reduction of the time between samples can be obtained by either reducing the inter-arrival time at a machine level or increasing the sampling frequency of the monitored products.

The formula for the average number of items between consecutive samples was determined by considering the relationship between the two performance measures. The formula reveals that the number of consecutive unsampled items depends on the volume fraction of the monitored product items and their corresponding sampling intervals. Therefore, whereas the time between samples is affected by the absolute values of the inter-arrival times at a machine level, the number of items between samples depends on a relative measure of the inter-arrival time. In particular, unlike the time between samples, it is not affected by the number of parallel machines in a station; however, it is affected by the presence of an unmonitored flow, which acts as a scale factor since it linearly varies the volume fractions of the different products.

The formulae for the average values were obtained using the results relative to all the stations in the segment; this means that they are valid no matter what the nature of the station is, whether a sampling or a non sampling station. However, the analogy between the stations is limited to the average values. In fact, the observation of the distribution of the number of consecutive unsampled items reveals that noticeable differences are present between the stations. The analysis focus hereinafter was exclusively addressed to the number of items between consecutive samples. The reason for this was mainly related with the fact that, as suggested by the results obtained, the number of items between consecutive samples represents a more global measure, since information about the unmonitored flow is also included in it.

The non-sampling station case was considered first. The regularity of the shape of the distribution for the different non-sampling stations and under different product flow conditions restricted the number of theoretical distributions suitable for modelling the number of items between samples. The nature of the events under investigation provided the needed support to state that the number of unsampled items between consecutive samples follows a geometric distribution with proportion p equal to the sampled fraction. In fact, the succession of sampled and unsampled items produced by a machine in a non-sampling station is a geometric process, in the sense that an item moving out of a machine could be a sample or not independently of the nature of the item previously processed by the same machine. The nature of the item moving out of a machine is exclusively based on the proportion of the population, that is, on the fraction of the sampled and unsampled items. This is immediately true when the item arrival and departure from a machine follows an exponential distribution, owing to its memory-less property. However, it was found that the assumption of an exponential inter-arrival time is not crucial for the validity of the results. The prediction accuracy proved very high; with respect to all the scenarios investigated the maximum absolute error was less than 99.6%, whereas the cumulative error range was always lower than 94%.

Given the nature of the formulae on which it is built, the presumable generality of the geometric model was tested against two scenarios with non-sampling stations characterised by different product flow conditions. The first scenario analysed the distribution of the number of unsampled items between consecutive samples in a non-sampling station, shared by two production segments, where samples could indifferently come from both the segments. The second case considered a station partially skipped by the global flow crossing the entire segment. In both cases the geometric model provided very accurate results. Finally, for a better understanding of the dynamics behind the generation of the distribution of the number of unsampled items, a basic simulation model, consisting of two stations and an intermediate buffer, was developed. Different elements of time related randomness were progressively introduced in the model. It was shown that the level of item sequence disorder and the randomness of the routing patterns between the machines of two successive stations are fundamental for the validity of the model, more than any particular hypothesis on the shape of the time input distributions.

The analysis of the sampling station case was split into three different parts, which correspond with the different product flow combinations considered in this research.

First, the case with one monitored product and an unmonitored flow was considered; then the analysis of the case with two monitored products without and with unmonitored flow followed.

For the first case, it was straightforward to find that the number of unsampled items follows a negative binomial distribution, with parameters p and s given by the monitored volume fraction and the sampling interval of the only monitored items, respectively. The prediction accuracy is still very high, with an absolute error less than 0.4%.

The challenge with the second case was to find a solution to the problem of generating a probability distribution by summing up deterministic values. After a few analyses, the enumeration of all the possible item sequences when a predetermined number of items of both the products were considered proved to be an effective approach. Due to its combinatorial nature, the approach developed is quite computationally expensive, so that only a limited range of sampling intervals could be investigated. However, it can predict quite well the distribution of the number of unsampled items, which for this case, presents a limited domain. The presence of a small but systematic error would require further investigation; however, it is very probable that this error is caused by the limited exploration of the whole population of the item sequences that was necessitated by the limited capabilities of the computer used during the analysis.

As regards the third case, the negative binomial distribution can provide an approximated prediction of the distribution of the number of unsampled items. The approximation depends on the fact that the average sampling interval of the two products is used as a parameter which is supposed to be deterministic. In effect, the global sampling interval is actually better described by the distribution of the number of unsampled monitored items, which is the distribution of the number of unsampled items when only monitored products are considered. This is the distribution obtained in the second case analysed. Hence, the prediction model for this last case analysed is obtained as a negative binomial distribution compound with the distribution obtained for the second case. The model is quite accurate; however, a variant developed outperforms it provided that a proper weighting factor is chosen. This optimal factor was found by means of a trial and error search, which is not a feasible approach when simulation results are not available. From analyses conducted on the robustness of the prediction model it emerged that the optimal factor is very likely connected with the standard deviation of the distribution of the number of unsampled items, which at the moment is still not predictable.

For the third contribution of this study, a few suggestions were presented on how to use the prediction models developed for the assessment of the quality risk associated with a sampling strategy. In summary, given a confidence level, the corresponding percentile of the distribution of the number of unsampled items between samples represents the quantification of the quality risk in terms of maximum number of items which might not be consecutively sampled (at that risk level). The prediction model to be considered for the risk assessment is obviously the one relative to the scenario which is observed in the station under investigation. It is worth noting that, due to its higher variability, the consideration of the geometric model will always provide more conservative estimates with respect to the other models developed. This observation is particularly useful when a sampling strategy satisfying that requirement has to be defined given an acceptable quality risk level. An iterative search allows the calculation of the minimum sampled fraction which guarantees the verification of the quality risk constraint in a non-sampling station. The same sampled fraction will also generate a full respect of the constraint in the sampling station of the segment. Based on the sampled fraction, a formula proposed allows the derivation of the sampling strategy parameters which satisfy the quality risk specifications. The use of the geometric model for this search is convenient for two reasons. First, the results based on it assure that the constraints will be respected everywhere in the segment, provided that the global flow is constant throughout the segment. Second, unlike the other models, the geometric one is characterised by only one parameter, so that the unknown variables are initially reduced to the minimum and no distribution prediction is needed.

Chapter VII

Conclusions and Future Work

7.1 Conclusions

The quality risk assessment in a complex manufacturing environment was the compelling reason for this research. The objective to turn random measures into predictable variables was its main challenge.

In this study, the quantification of the quality risk passes through three fundamental steps

1. The investigation of the mechanisms governing the effects of the implementation of a particular sampling strategy behaviour in a system characterised by sequence disorder effect and multiple stream effect;
2. The development of prediction models for a quality risk related performance measure;
3. The definition of a procedure for predicting the quality risk.

The reasons for this progression can be found in the original concept of the quality risk related measure. It should have expressed the maximum exposition of any machine in the system to the risk of not being monitored. This was deliberately quantified in terms of the maximum number of consecutive unsampled items produced by that machine for a given confidence level. The choice to relate this risk with a confidence level depends on both the concept of risk, which does not suit any definite notion, and the random nature of the number of consecutive unsampled items, which did not allow for the quantification of the risk in terms of an absolute maximum measure. The necessity for developing prediction models for the distribution of the number of unsampled items between samples is a direct consequence of this choice. This is the reason why this study could not be considered complete when the formulae for the prediction of the average value of the performance measures were developed. They are able to express in an effective, comprehensive way what the analysis initially conducted, based on a simulation approach, had revealed in a more fractional fashion. They are fundamental for this work since

1. they support a general evaluation of the sampling strategy effectiveness, which is not strictly related with the risk;
2. most of the prediction models developed for the distribution of the number of unsampled items uses the formulae as an input.

The initial analysis represents the solid foundation of this work. The relationship between some control parameters and the performance measures were explored, the discernment of the most affecting parameters was made and an initial understanding of the interventions useful to reduce the quality risk was possible.

Apart from the quantification of the quality risk, the most important achievement of this work is the possibility of developing a sampling strategy able to keep the quality risk under the desired threshold level. The most important findings are

1. the average time between samples depends on the inter-arrival time at a machine level and the sampling intervals of the monitored products crossing a station;
2. in all the stations, the average number of unsampled items is inversely proportional to the sampled fraction;
3. the number of unsampled items between consecutive samples in a non-sampling station follows a geometric distribution. This is valid for any material flow scenarios;
4. the distribution of the number of unsampled items in a sampling station can be predicted using more complicated models which depend on the material flow combinations;

5. the geometric model provides a conservative estimate of the quality risk in any station.

7.2 Recommendations for future work

The list of the principal elements which are worthy to investigate in the future mainly derives from the set of issues which have been neglected during the development of the thesis. Initially, these issues were deemed not immediately fundamental for the development of reasonably well approximated solutions. In fact, the results obtained are characterised by a very high accuracy, as regards the prediction models, and an immediate simplicity relative to the procedures for the quality risk assessment. However, going back to those issues and trying to pay more attention to them could surely prove beneficial for the analysis rigours and confer on the solutions proposed a more comprehensive frame.

a) The time between samples should be analysed in terms of its distribution. The lack of a prediction model for the distribution of the time between samples represents the main negligence in this thesis. If for the industrial environments characterised by the production of discrete items, the availability of a time related quality risk measure merely represents an alternative option to the use of the number of items, the assessment of the quality risk from a time viewpoint becomes fundamental for the continuous production environments, where the lack of discrete items makes the application of the results obtained so far difficult. The continuous nature represents a very challenging element for the development of the models.
b) The systematic nature of the prediction error pattern for the distribution of the number of unsampled monitored items should be analysed and a correction to the prediction model should be proposed, whenever, unlike how suspected, the systematic error is not caused by the limited exploration of the item sequence space.
c) The applicability of the prediction model of the distribution of the number of unsampled items developed for scenarios characterised by two monitored products should be tested for scenarios involving more than two monitored products. The

availability of more powerful calculators should allow the analysis of more complex scenarios.

d) The optimal weighting factor used in the correction variant of the prediction model for the distribution of the number of unsampled items for scenarios characterised by two monitored product types and an unmonitored flow should be predicted by means of analytical formula. More investigations are needed to find the parameters which mostly affect it. As suggested by the results found, these might include the standard deviation of the final distribution, which would be advisable to predict independently of its actual usefulness to the weighting factor definition.

e) The analysis of the case when in a sampling station sampled items coming from elsewhere in the segment merges with the samples chosen in the station might be challenging. In fact, in this case, a deterministic and a random element are both associated with the monitored flow. Presumably, it will be difficult to find straightforward solutions in the commonly known distributions and an enumeration approach could prove necessary.

f) The investigation of the robustness of the prediction models to slight variations of the sampling strategy dynamics could also be interesting. For example, the analysis of the system reaction to sampling performed on a random basis in the sampling station represents a possible variant to be considered.

g) Exploring the applicability of enumerating techniques to a more general class of problems and developing approaches to reduce their computational complexity could prove interesting.

Bibliography

[1] Parasuraman, A., Zeithaml, V.A., Berry, L.L., (1985), "A conceptual model of service quality and its implications for future research", *The Journal of Marketing*, Vol.49 (4), pp. 41-50.

[2] Reeves, C.A., Bednar, D.A., (1994), "Defining quality: alternatives and implications", *Academy of Management Review*, Vol.19 (3), pp. 419-445.

[3] Ferme, W.D., (1995), "Design for quality", *Proceedings of the 1995 IEEE Annual Internaional Engineering Management Conference, 1995.'Global Engineering Management: Emerging Trends in the Asia* Pacific', pp. 229-234.

[4] Schoemig, A.K., (1999), "On the corrupting influence of variability in semiconductor manufacturing", "On the corrupting influence of variability in semiconductor manufacturing", *Proceedings of the 31st conference on Winter simulation: Simulation---a bridge to the future-Volume 1,* ACM New York, NY, USA, pp. 837-842.

[5] Taylor, G.D., Heragu, S.S., (1999), "A comparison of mean reduction versus variance reduction in processing times in flow shops", *International Journal of Production Research,* Vol.37 (9), pp. 1919-1934.

[6] Menipaz, E., (1978), "A taxonomy of economically based quality control procedures", *International Journal of Production Research,* Vol.16 (2), pp. 153-167.

[7] Chan, K.K., Spedding, T.A., (2001), "On-line optimization of quality in a manufacturing system", *International Journal of Production Research,* Vol.39 (6), pp. 1127-1145.

[8] Oppermann, M., Sauer, W., Wohlrabe, H., Zerna, T., (2003), "New quality cost models to optimize inspection strategies", *IEEE Transactions on Electronics Packaging Manufacturing,* Vol.26 (4), pp. 328-337.

[9] Penn, M., Raviv, T., (2007), "Optimizing the quality control station configuration", *Naval Research Logistics,* Vol.54 (3), pp. 301-314.

[10] Fan, C.M., Guo, R.S., Chang, S.C., Wei, C.S., (2000), "SHEWMA: an end-of-line SPC scheme using wafer acceptance test data", *IEEE Transactions on Semiconductor Manufacturing,* Vol.13 (3), pp. 344-358.

[11] Jin, M., Li, Y., Tsung, F., (2010), "Chart allocation strategy for serial-parallel multistage manufacturing processes", *IIE Transactions,* Vol.42 (8), pp. 577-588.

[12] Huang, Q., Shi, J., (2004), "Stream of variation modeling and analysis of serial-parallel multistage manufacturing systems", *Journal of Manufacturing Science and Engineering,* Vol.126 (3), pp. 611-619.

[13] Zeng, L., Zhou, S., (2007), "Inferring the interactions in complex manufacturing processes using graphical models", *Technometrics*, Vol.49 (4), pp. 373-381.

[14] Guo, R. S., Tsai, C.K., Lee, G.H., Chang, S.C., (1996), "Intelligent process diagnosis based on end-of-line electrical test data", *Proceeding of the 1996 Electronics Manufacturing Technology Symposium*, pp. 347-354.

[15] Montijn-Dorgelo, F.N.H., ter Horst, H.J. (1988), "Expert System for Test Structure Data Interpretation", *Proceedings of the 1998 IEEE International Conference on Microelectronic Test Structures, 1988. ICMTS*, pp. 169-173.

[16] Pak, J., Kittler, R., Hopper, B., Wen, P., (1997), "Advanced methods for analysis of wafer-to-wafer yield variation", *IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop, 1997*, pp. 62-66.

[17] Fan, J. M., Guo, R.S., Chang, S.C., Lee, J.H., (1996), "Abnormal Trend Detection of Sequence-Disordered Data Using EWMA Method", *1996 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, pp. 169-174.

[18] Fan, C.M., Guo, R.S., Chang, S.C., (1997), "EWMA/SD: an End-of-Line SPC Scheme to Monitor Sequence-Disordered Data", *2nd International Workshop on Statistical Metrology*, Vol.8, pp. 66-69.

[19] Fan, C. M., Chang, S.C., Guo, R.S., Kung, H.H., You, J.K., Chen, H.P., Lin, S., Wei, J., (1999), "SHEWMAC: an end-of-line SPC scheme for joint monitoring of process mean and variance", *Proceedings of the 1999 IEEE International Symposium on Semiconductor Manufacturing Conference*, pp. 233-236.

[20] Fan, C. M., Chang, S.C., Guo, R.S., Kung, H.H., You, J.K., Chen, H.P., Lin, S., Wei, J., (2000), "SHEWMAC: an end-of-line SPC scheme via exponentially weighted moving statistics", *Semiconductor Manufacturing Technology Workshop, 2000*, pp. 268-278.

[21] Inman, R., Blumenfeld, D.E., Huang, N., Li, J., (2003), "Designing production systems for quality: research opportunities from an automotive industry perspective", *International Journal of Production Research*, Vol.41 (9), pp. 1953-1971.

[22] Kim, J., Gershwin, S.B., (2005), "Integrated quality and quantity modeling of a production line", *OR Spectrum*, Vol.27 (2), pp. 287-314.

[23] Kim, J., Gershwin, S.B., (2008), "Analysis of long flow lines with quality and operational failures", *IIE Transactions*, Vol.40 (3), pp. 284-296.

[24] Colledani, M., Tolio, T., (2006), "Impact of quality control on production system performance", *CIRP Annals-Manufacturing Technology*, Vol.55 (1), pp. 453-456.

[25] Colledani, M., Tolio, T., (2009), "Performance evaluation of production systems monitored by statistical process control and off-line inspections", *International Journal of Production Economics*, Vol.120 (2), pp. 348-367.

[26] Schick, I.C., Gershwin, S.B., Kim, J., (2005), "Quality/Quantity modeling and analysis of production lines subject to uncertainty", *Final report presented to General Motors research and development center*.

[27] Erdinca, O., Yeowb, P.H.P., (2010), "Proving external validity of ergonomics and quality relationship through review of real-world case studies", *International Journal of Production Research*, iFirst, pp. 1-14.

[28] Li, J., Blumenfeld, D.E., Marin, S.P., (2008), "Production system design for quality robustness", *IIE Transactions*, Vol.40 (3), pp. 162-176.

[29] Blumenfeld, D.E., Inman, R.R., (2009), "Impact of absenteeism on assembly line quality and throughput", *Production and Operations Management,* Vol.18 (3), pp. 333-343.

[30] Huang, N., Inman, R., (2010), "Product quality and plant build complexity", *International Journal of Production Research,* Vol.48 (11), pp. 3105-3128.

[31] Owen, J., Blumenfeld, D., (2008), "Effects of operating speed on production quality and throughput", *International Journal of Production Research,* Vol.46 (24), pp. 7039-7056.

[32] Zeithaml, V.A., (1988), "Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence", *The Journal of Marketing,* Vol.52 (3), pp. 2-22.

[33] Sethi, R., (2000), "New product quality and product development teams", *The Journal of Marketing,* Vol.64 (2), pp. 1-14.

[34] Montgomery, D.C., (2008), *Introduction to statistical process control,* (5th ed.), John Wiley & Sons, Inc.

[35] Taylor, F.W., (1911), *The principles of scientific management,* Adamant Media Corporation.

[36] Doty, L.A., (1996), *Statistical process control,* Industrial Press Inc.

[37] Shewhart, W.A., (1980), *Economic control of quality of manufactured product,* American Society for Quality.

[38] Duncan, A.J., (1956), "The Economic Design of barX Charts used to Maintain Current Control of a Process", *Journal of the American Statistical Association,* Vol.51 (274), pp. 228-242.

[39] Spanos, C.J., (1992), "Statistical process control in semiconductor manufacturing", *IEEE Proceedings,* Vol.80 (6), pp. 819-830.

[40] Dodge, H.F., Romig, H.G., (1959), *Sampling inspection tables,* Wiley New York, NY.

[41] Lindsay, G.F., Bishop, A.B., (1964), "Allocation of Screening Inspection Effort-A Dynamic-Programming Approach", *Management Science,* Vol.10 (2), pp. 342-352.

[42] White, L.S., (1969), "Shortest route models for the allocation of inspection effort on a production line", *Management Science,* Vol.15 (5), pp. 249-259.

[43] Britney, R.R., (1972), "Optimal screening plans for non-serial production systems", *Management Science,* Vol.18 (9), pp. 550-559.

[44] Raz, T., (1986), "A survey of models for allocating inspection effort in multistage production systems.", *Journal of Quality Technology,* Vol.18 (4), pp. 239-247.

[45] Beightler, C., Mitten, L., (1964), "Design of an optimal sequence of interrelated sampling plans", *Journal of the American Statistical Association,* Vol.59 (305), pp. 96-104.

[46] Woo, W.K., Metcalfe, J.E., (1972), "Optimal allocation of inspection effort in multistage manufacturing processes", *Western Electric Engineer,* Vol.16 (3), pp. 8-16.

[47] Van Volsem, S., Dullaert, W., Van Landeghem, H., (2007), "An Evolutionary Algorithm and discrete event simulation for optimizing inspection strategies for multi-stage processes", *European Journal of Operations Research,* Vol.179 (3), pp. 621-633.

[48] Sarhangian, V., Vaghefi, A., Eskandari, H., Ardakani, M.K., (2008), "Optimizing inspection strategies for multi-stage manufacturing processes using simulation optimization", *Proceedings of the 40th Conference on Winter Simulation*, pp. 1974-1980.

[49] Vaghefi, A., Sarhangian, V., (2009), "Contribution of simulation to the optimization of inspection plans for multi-stage manufacturing systems", *Computers and Industrial Engineering*, Vol. 57, pp. 1226-1234.

[50] Heredia-Langner, A., Montgomery, D.C., Carlyle, W.M., (2002), "Solving a multistage partial inspection problem using genetic algorithms", *International Journal of Production Research,* Vol.40 (8), pp. 1923-1940.

[51] Juran, J.M., (1995), "World War II and the quality movement", in *Quality in Higher Education,* pp. 65.

[52] Arnheiter, E.D., Maleyeff, J., (2005), "The integration of lean management and Six Sigma", *The TQM Magazine,* Vol.17 (1), pp. 5-18.

[53] Girshick, M.A., Rubin, H., (1952), "A Bayes approach to a quality control model", *The Annals of mathematical statistics,* Vol.23 (1), pp. 114-125.

[54] Giakatis, G., Enkawa, T., Washitani, K., (2001), "Hidden quality costs and the distinction between quality cost and quality loss", *Total Quality Management & Business Excellence,* Vol.12 (2), pp. 179-190.

[55] Sandoval-Chávez, D.A., Beruvides, M.G., (1998), "Using opportunity costs to determine the cost of quality: A case study in a continuous-process industry", *The Engineering Economist,* Vol.43 (2), pp. 107-124.

[56] Chang, S.J., Park, Y.H., Park, E.H., (1996), "Quality costs in manufacturing systems", *Computers and Industrial Engineering,* Vol.31 (1), pp. 115-118.

[57] Freeman, J.M., (2008), "The case for quality costing simulation", *The TQM Journal,* Vol.20 (5), pp. 476-487.

[58] Chien, C. F., Hsu, S.H., Peng, S., Wu, C.H., (2000), "A cost-based heuristic for statistically determining sampling frequency in a wafer fab", *2000 Semiconductor Manufacturing Technology Workshop,* pp. 217-229.

[59] Tapiero, C.S., Hsu, L.F., (1987), "Quality control of the M/M/1 queue", *International Journal of Production Research,* Vol.25 (3), pp. 447-455.

[60] Nurani, R.K., Akella, R., Strojwas, A.J., (1996), "In-line defect sampling methodology in yield management: anintegrated framework", *IEEE Transactions on Semiconductor Manufacturing,* Vol.9 (4), pp. 506-517.

[61] Young, H.A., (1963), "On the optimum location of checking stations", *Operations Research,* Vol.11 (5), pp. 721-731.

[62] Hsu, J.I.S., (1984), "Hybrid inspection system for the multistage production process", *International Journal of Production Research,* Vol.22, pp. 63-69.

[63] Alfares, H., (1999), "A simulation model for determining inspection frequency", *Computers and Industrial Engineering,* Vol.36 (3), pp. 685-696.

[64] Engi˙ n, A.B., (2008), "Determination of optimum economic inspection by economic control chart design and by machine efficiency estimation: An application in weaving industry", *Simulation Modelling Practice and Theory,* Vol.16 (1), pp. 147-170.

[65] Ng, W.C., Hui, Y.V., (1997), "Economic design of a complete inspection plan with interactive quality improvement", *European Journal of Operations Research,* Vol.96 (1), pp. 122-129.

[66] Rau, H., Cho, K.H., (2009), "Genetic algorithm modeling for the inspection allocation in reentrant production systems", *Expert Systems with Applications*, Vol. 36 (8), pp. 11287-11295.

[67] Oppermann, M., Sauer, W., Wohlrabe, H, Zerna, T., (2001), "Optimization of inspection strategies by use of quality cost models", *Proceedings of the 51st Electronic Components and Technology Conference, 2001,* pp. 1321-1325.

[68] McIntyre, M., Nurani, R.K., Akella, R., (1996), "Key considerations in the development of defect samplingmethodologies", *Proceedings of the IEEE/SEMI 1996 Advanced Semiconductor Manufacturing Conference and Workshop, 1996, ASMC 96,* pp. 81-85.

[69] Kuo, W.W., Akella, R., Fletcher, D., (1997), "Adaptive sampling for effective multi-layer defect monitoring", *Proceedings of the 1997 IEEE International Symposium on Semiconductor Manufacturing Conference,* pp. D1-D4.

[70] Hall, G. D. R., Young, R., Dunne, M., Muro, M., (2008), "A quality-cost model of in-line inspections for excursion detection and reduction", *Proceedings of the IEEE/SEMI 2008 Advanced Semiconductor Manufacturing Conference*, pp. 273-277.

[71] Lin, L., Chen, S., Bousetta, A., Yang, R., Liao, J., (2005), "Improving defect capture rate for excursion detection", *Semicond International,* Vol.28 (9), pp. 55-58.

[72] Jang, W., Wang, E.H., Akella, R., (2000), "Economic in-line inspection sampling strategy with learning effects", *Int J Prod Res,* Vol.38 (18), pp. 4811-4822.

[73] Duffuaa, S.O., (1996), "Impact of inspection errors on the performance measures of a complete repeat inspection plan", *International Journal of Production Research*, Vol.34, pp. 2035-2049.

[74] Duffuaa, S.O., Khan, M., (2005), "Impact of Inspection Errors on the Performance of Measure of a General Repeat Inspection Plan", *International Journal of Production Research*, Vol.43, pp. 4945-4967.

[75] Elshafei, M., Khan, M., Duffuaa, S.O., (2006), "Repeat inspection planning using dynamic programming", *International Journal of Production Research*, Vol.44 (2), pp. 257-270.

[76] Raouf, A., Jain, J.K., Sathe, P.T., (1983), "A cost-minimization model for multicharacteristic component inspection", *IIE Transactions,* Vol.15 (3), pp. 187-194.

[77] Rau, H., Chu, Y.H., (2005), "Inspection allocation planning with two types of workstation: WVD and WAD", *The International Journal of Advanced Manufacturing Technology,* Vol.25 (9), pp. 947-953.

[78] Sheu, S.H., Chen, Y.C., Wang, W.Y., Shin, N.H., (2003), "Economic optimization of off-line inspection with inspection errors", *The Journal of the Operational Research Society,* pp. 888-895.

[79] Ballou, D.P., Pazer, H.L., (1982), "The impact of inspector fallibility on the inspection policy in serial production systems", *Management Science,* Vol.28 (4), pp. 387-399.

[80] Lee, J., Unnikrishnan, S., (1998), "Planning quality inspection operations in multistage manufacturing systems with inspection errors", *International Journal of Production Research,* Vol.36 (1), pp. 141-156.

[81] Bendavid, I., Herer, Y.T., (2009), "Economic optimization of off-line inspection in a process that also produces non-conforming units when in control and conforming units when out of control", *European Journal of Operational Research.,* Vol.195 (1), pp. 139-155.

[82] Wang, C.H., (2007), "Economic off-line quality control strategy with two types of inspection errors", *European Journal of Operational Research,* Vol.179 (1), pp. 132-147.

[83] Sakurai, K., Fujii, S., Morita, S., (2001), "A basic study on defect inspection method for ASIC manufacturing line (First Report) Inspection parameter dependence on yield and cleaning cycle (Second Report)", *IEE Transactions Japan,* Vol.121, pp. 231.

[84] Lee, S.B., Lee, T.Y., Liao, J., Chang, Y.C., (2003), "A capacity-dependence dynamic sampling strategy", *2003 IEEE International Symposium on Semiconductor Manufacturing,* pp. 312-314.

[85] Chen, K. D., Akella, R., Emami, I., McIntyre, M., (1997), "Effective multi-stage test equipment capacity allocation forsemiconductor fabrication yield enhancement", *Proceedings of the 1997 IEEE International Symposium on Semiconductor Manufacturing Conference,* pp. P73-P75.

[86] Kouikoglou, V.S., Phillis, Y.A., (2002), "Design of product specifications and control policies in a single-stage production system", *IIE Transactions,* Vol.34 (7), pp. 591-600.

[87] Chen, F.F., Everett, E.A., (1991), "The impact of flexible manufacturing systems on productivity andquality", *IEEE Transactions on Engineering Management,* Vol.38 (1), pp. 33-45.

[88] Mapes, J., New, C., Szwejczewski, M., (1997), "Performance trade-offs in manufacturing plants", *International Journal of Operations and Production Management,* Vol.17, pp. 1020-1033.

[89] Gershwin, S.B., Schick, I.C., (2007), "A taxonomy of quality/quantity issues in manufacturing systems", *Proceedings of the Sixth Conference on the Analysis of Manufacturing Systems, Lunteren, the Netherlands,* pp. 1–6.

[90] Ioannidis, S., Kouikoglou, V.S., Phillis, Y.A., (2004), "Coordinating quality, production and sales in manufacturing systems", *International Journal of Production Research,* Vol.42 (18), pp. 3947-3956.

[91] Carcano, O.E., Portioli-Staudacher, A., (2006), "Integrating inspection-policy design in assembly-line balancing", *International Journal of Production Research,* Vol.44 (18), pp. 4081-4104.

[92] Yu, H.F., Yu, W.C., (2008), "The optimal production and quality policy for the vendor in a trade", *International Journal of Production Research,* Vol.46 (15), pp. 4135-4153.

[93] Chakraborty, T., Giri, B.C., Chaudhuri, K.S., (2008), "Production lot sizing with process deterioration and machine breakdown", *European Journal of Operational Research,* Vol.185 (2), pp. 606-618.

[94] Chakraborty, T., Giri, B.C., Chaudhuri, K.S., (2009), "Production lot sizing with process deterioration and machine breakdown under inspection schedule", *Omega,* Vol.37 (2), pp. 257-271.

[95] Wang, C.H., Sheu, S.H., (2001), "Simultaneous determination of the optimal production–inventory and product inspection policies for a deteriorating production system", *Computers and Operations Research,* Vol.28 (11), pp. 1093-1110.

[96] Sarker, B.R., Jamal, A.M.M., Mondal, S., (2008), "Optimal batch sizing in a multi-stage production system with rework consideration", *European Journal of Operational Research,* Vol.184 (3), pp. 915-929.

[97] Wang, C.H., (2005), "Integrated production and product inspection policy for a deteriorating production system", *International Journal of Production Economics,* Vol.95 (1), pp. 123-134.

[98] Huang, L.H., (2009), "Labour assignment and workload balance evaluation for a production line", *Estonian Journal of Engineering,* Vol.15 (1), pp. 34-47.

[99] Boring, R.L., Gertman, D.I., (2004), "Human error and available time in SPAR-H", *Workshop on Temporal Aspects of Work for HCI, CHI 2004*, pp. 1-4.

[100] ElMaraghy, W.H., Nada, O.A., ElMaraghy, H.A., (2008), "Quality prediction for reconfigurable manufacturing systems via human error modelling", *International Journal of Computer Integrated Manufacturing,* Vol.21 (5), pp. 584-598.

[101] Ackoff, R.L., (1994), "Systems thinking and thinking systems", *System Dynamics Review,* Vol.10 (2-3), pp. 175-188.

[102] Chan, L.Y., Xie, M., Goh, T.N., (2000), "Cumulative quantity control charts for monitoring production processes", *International Journal of Production Research,* Vol.38 (2), pp. 397-408.

[103] Xie, M., Goh, T.N., Ranjan, P., (2002), "Some effective control chart procedures for reliability monitoring", *Reliability Engineering and System Safety,* Vol.77 (2), pp. 143-150.

[104] Chan, L., Lin, D.K.L., Xie, M., Goh, T.N., (2002), "Cumulative probability control charts for geometric and exponential process characteristics", *International Journal of Production Research*, Vol.40 (1), pp. 133-150.

[105] Zhang, C.W., Xie, M., Goh, T.N., (2005), "Economic design of exponential charts for time between events monitoring", *International Journal of Production Research,* Vol.43 (23), pp. 5019-5032.

[106] Shamsuzzaman, M., Xie, M., Goh, T.N., Zhang, H., (2009), "Integrated control chart system for time-between-events monitoring in a multistage manufacturing system", *The International Journal of Advanced Manufacturing Technology,* Vol.40 (3), pp. 373-381.

[107] Veatch, M.H., (2000), "Inspection strategies for multistage production systems with time-varying quality", *International Journal of Production Research,* Vol.38 (4), pp. 837-853.

[108] Jin, M., Tsung, F., (2009), "A chart allocation strategy for multistage processes", *IIE Transactions,* Vol.41 (9), pp. 790-803.

[109] Lam, Y.C., Shamsuzzaman, M., Zhang, S., Wu, Z., (2005), "Integrated control chart system—optimization of sample sizes, sampling intervals and control limits", *International Journal of Production Research,* Vol.43 (3), pp. 563-582.

[110] Wu, Z., Shamsuzzaman, M., (2005), "Design and application of integrated control charts for monitoring process mean and variance", *Journal of Manufacturing Systems,* Vol.24 (4), pp. 302-314.

[111] Law, A.M., Kelton, W.D., (1991), "Simulation modelling and analysis", *New York et al*.

[112] Tolio, T., Matta, A., Jovane, F., (1998), "A method for performance evaluation of automated flow lines", *CIRP Annals-Manufacturing Technology*, Vol.47 (1), pp. 373-376.

[113] Delp, D., Si, J., Fowler, J.W., (2006), "The development of the complete X-factor contribution measurement for improving cycle time and cycle time variability", *IEEE Transactions on Semiconductor Manufacturing,* Vol.19 (3), pp. 352-362.

[114] Barney, M., (2002), "Motorola's second generation", *Six Sigma Forum Magazine,*Vol.1, pp. 13-16.

[115] Fan, C. M., Guo, R.S., Chen, A., Hsu, K.C., Wei, C.S., (2001), "Data mining and fault diagnosis based on wafer acceptance test dataand in-line manufacturing data", *2001 IEEE International Semiconductor Manufacturing Symposium,* pp. 171-174.

[116] Mortell, R.R., Runger, G.C., (1995), "Statistical process control of multiple stream processes", *Journal of Quality Technology*, Vol.27 (1), pp. 1-12.

[117] Liu, X., MacKay, R.J., Steiner, S., (2008), "Monitoring Multiple Stream Processes", *Quality Engineering,* Vol.20 (3), pp. 296-308.

[118] Meneces, N.S., Olivera, S.A., Saccone, C.D., Tessore, J., (2008), "Statistical Control of Multiple-Stream Processes: A Shewhart Control Chart for Each Stream", *Quality Engineering*, Vol.20 (2), pp. 185-194.

[119] Lanning, J.W., Montgomery, D.C., Runger, G.C., (2002), "Monitoring a Multiple Stream Filling Operation Using Fractional Samples", *Quality Engineering,* Vol.15 (2), pp. 183-195.

[120] Schmidt, J.W., Taylor, R.E., (1970), *Simulation and analysis of industrial systems,* RD Irwin.

[121] Shannon, R.E., (1992), "Introduction to simulation", *Proceedings of the 24th conference on Winter simulation,* ACM New York, NY, USA, pp. 65-73.

[122] Shannon, R.E., (1998), "Introduction to the art and science of simulation", *Proceedings of the 1998 conference on Winter simulation,* ACM New York, NY, USA, pp. 7-14.

[123] Sargent, R.G., (2005), "Verification and validation of simulation models", "Verification and validation of simulation models", *Proceedings of the 37th conference on Winter simulation,* pp. 157-169.

[124] Robinson, S., (2005), "Discrete-event simulation: from the pioneers to the present, what next?", *Journal of the Operational Research Society,* Vol. 56(6), pp. 619-629.

[125] April, J., Glover, F., Kelly, J.P., Laguna, M., (2003), "Simulation-based optimization: practical introduction to simulation optimization", *Proceedings of the 35th conference on Winter simulation: driving innovation,* pp. 71-78.

[126] Roy, R., (1998), "Scheduling and control, performance measures and discrete event simulation", *Journal of the Operational Research Society,* Vol. 49(2), pp. 151-156.

[127] Goel, S., Strong, D.R., Richards, N., Goel, N.C., (2002), "A simulation-based method for the process to allow continuous tracking of quality, cost, and time", *Simulation,* Vol.78 (5), pp. 330.

[128] Neubauer, A.S., (1997), "The EWMA control chart: properties and comparison with other quality-control procedures by computer simulation", *Clinical Chemistry,* Vol.43 (4), pp. 594-601.

[129] Flowers, A.D., Cole, J.R., (1985), "An application of computer simulation to quality control in manufacturing", *IIE transactions,* Vol.17 (3), pp. 277-283.

[130] Gardner, L.L., Grant, M.E., Rolston, L.J., (1995), "Using simulation to assess costs of quality", *Proceedings of the 27th conference on Winter simulation*, pp. 945-951.

[131] Freeman, J.M., (1995), "Estimating quality costs", *Journal of the Operational Research Society,* Vol.46 (6), pp. 675-686.

[132] De Ruyter, A.S., Cardew-Hall, M.J., Hodgson, P.D., (2002), "Estimating quality costs in an automotive stamping plant through the use of simulation", *International Journal of Production Research,* Vol.40 (15), pp. 3835-3848.

[133] Visawan, D., Tannock, J., (2004), "Simulation of the economics of quality improvement in manufacturing: A case study from the Thai automotive industry", *International Journal of Quality & Reliability Management,* Vol.21 (6), pp. 638-654.

[134] Lee, Y.H., Han, C.S., Jung, J.W., (2006), "Using capacity simulation for quality goal setting", *System Modelling and Simulation: Theory and Applications. 3rd Asian Simulation Conference, AsiaSim 2004*, pp. 31-39.

[135] Dodge, H.F., (1955), "Estimating quality costs in an automotive stamping plant through the use of simulation", *Industrial Quality Control,* Vol.11 (15), pp. 3-5.

[136] Lucas, J.M., Saccucci, M.S., (1990), "Exponentially weighted moving average control schemes: properties and enhancements", *Technometrics,* Vol.32 (1), pp. 1-12.

[137] Yang, Z., Xie, M., Kuralmani, V., Tsui, K.L., (2002), "On the performance of geometric charts with estimated control limits", *Journal of quality technology,* Vol.34 (4), pp. 448-458.

[138] Jones, L.A., Champ, C.W., Rigdon, S.E., (2004), "The run length distribution of the CUSUM with estimated parameters", *Journal of Quality Technology,* Vol.36 (1), pp. 95-108.

[139] Ignall, E.J., Kolesar, P., Walker, W.E., (1978), "Using simulation to develop and validate analytic models: some case studies", *Operations Research,* Vol.26 (2), pp. 237-253.

[140] Shanthikumar, J.G., Sargent, R.G., (1983), "A unifying view of hybrid simulation/analytic models and modeling", *Operations Research,* Vol.31 (6), pp. 1030-1052.

[141] Nyhuis, P., von Cieminski, G., Fischer, A., (2005), "Applying simulation and analytical models for logistic performance prediction", *CIRP Annals-Manufacturing Technology,* Vol.54 (1), pp. 417-422.

[142] Wang, W., (2003), "Modelling condition monitoring intervals: A hybrid of simulation and analytical approaches", *Journal of the Operational Research Society,* Vol.54 (3), pp. 273-282.

[143] Byrne, M., Bakir, M., (1999), "Production planning using a hybrid simulation–analytical approach", *International Journal of Production Economics,* Vol.59 (1-3), pp. 305-311.

[144] Byrne, M., Hossain, M., (2005), "Production planning: An improved hybrid approach", *International Journal of Production Economics,* Vol.93, pp. 225-229.

[145] Kim, B., Kim, S., (2001), "Extended model for a hybrid production planning approach", *International Journal of Production Economics,,* Vol.73 (2), pp. 165-173.

[146] Papoulis, A., Pillai, S.U., (2002), *Probability, random variables and stochastic processes,* McGraw-Hill Education (India) Pvt Ltd.

[147] Meyer, P.L., (1980), *Introductory probability and statistical applications,* Addison-Wesley Reading, Mass.

[148] Savage, I.R., (1957), "Nonparametric statistics", *Journal of the American Statistical Association,* Vol.52 (279), pp. 331-344.

[149] Avis, K.E., (1995), *Sterile pharmaceutical products: process engineering applications*, CRC,.

[150] Kuo, Y., Yang, T., (2006), "A case study on the operator allocation decision for TFT-LCD inspection and packaging process", *Journal of Manufacturing Technology Management*, Vol.17 (3), pp. 363-375.

[151] Landau, K., Peters, H., (2006), "Ergonomic demands in automotive component inspection tasks", *Occupational Ergonomics,* Vol.6 (2), pp. 95-105.

[152] Narahari, Y., Khan, L.M., (1995), "Modeling re-entrant manufacturing systems with inspections", *Proceedings of the IEEE International Conference on Robotics and Automation, Japan,* Vol.2, pp. 367-378.