Dauzère-Pérès, S., & Hassoun, M. (2020). On the importance of variability when managing metrology capacity. *European Journal of Operational Research*, *282*(1), 267–276. https://doi.org/10.1016/j.ejor.2019.09.014

On the importance of variability when managing metrology capacity

Stéphane Dauzère-Pérès[1,2] Michael Hassoun[3]

[1]Mines Saint-Etienne, Univ Clermont Auvergne
CNRS, UMR 6158 LIMOS
CMP, Department of Manufacturing Sciences and Logistics
Gardanne, France
E-mail: dauzere-peres@emse.fr

[2]Department of Accounting, Auditing and Business Analytics
BI Norwegian Business School
Oslo, Norway

[3]Department of Industrial Engineering and Management
Ariel University
Ariel, Israel
E-mail: michaelh@ariel.ac.il

Abstract

In-line quality control is a crucial and increasingly constraining activity, in particular in high technology manufacturing. In this paper, we study a single metrology tool assigned to control the production quality of multiple heterogeneous machines. We introduce, model and study the tradeoff between the quality loss resulting from the sampling policy, and the quality loss induced by delays in the metrology queue. An iterative approach is proposed to optimize sampling periods using the solution of a relaxed problem which assumes full synchronization between production and metrology, and which has been previously formalized and solved. Based on computational and simulation results, and a prediction model, the paper ends with recommendations to better manage metrology capacity utilization under various levels of variability.

Keywords: Manufacturing, Metrology, Variability, Discrete Optimization, Queueing Systems

1 Introduction

For some time now, both operation managers and researchers have identified the need to integrate metrology (also called inspection, control or measurement) operations, usually dictated by quality requirements, to their operational and economical environment (see Goyal et al. (1993) for a review). For instance, Gilenson et al. (2015) propose to balance between quality and throughput. Several examples of economic optimization of Statistical Process Control (SPC) can be found in Sultan and Rahim (1997). Bouslah et al. (2016) study the aspects of controlling related to preventive maintenance. In all these domains, the metrology capacity and the queue forming at the metrology tool (any facility or machine performing inspection operations) are seldom considered. A counterexample may be found in Tang (1991) that solves a non-linear integer program to determine, among other variables, the number and location of inspection tools in a transfer line modeled as a queueing network. We will make use of similar ideas in this work, where we aim at modeling and proposing an approach that takes into account the impact of the queue in metrology on the risk when production continues while the product to be measured waits in front of the metrology tool.

The relative lack of interest for quality-inspection capacitated problems may be explained by the traditionally low cost of and limited space taken by metrology tools. This state of affairs is quickly changing and various manufacturing sectors increasingly depend on complex, expensive and large metrology tools to control their process. The most prominent example is certainly in semiconductor manufacturing, where the seemingly never-ending race for smaller pitch in technology challenges the ability of quality control to follow up. The frequency and the sensibility of controlling production processes, and the cost of equipment have turned metrology into a high utilization area, and managers are struggling to achieve what they see as the required level of control (see Colledani and Tolio (2011), Bettayeb et al. (2012), Nduhura-Munga et al. (2012), Lee et al. (2003), Shanoun et al. (2011) and the review in Nduhura-Munga et al. (2013)).

In contrast, the literature on management of congestion on production machines is abundant. Several researchers have proposed to control the level of utilization of production systems in order to coordinate between lead times, release times and service levels. For example, Hendry et al. (1998) and Kingsman and Hendry (2002) show how Work Load Control (WLC) planning systems are beneficial to lead time and thus to service level in make-to-order plants. Zäpfel and Missbauer (1993) discuss the pertinence of aggregate models to determine the parameters of a WLC planning system. In another vein, Orcun et al. (2009) tackle the non linear dependence between workload and lead times by using clearing functions, and provide a release schedule designed to reach a target service level. Again, to the best of our knowledge, while widely accepted for production, these concepts have not been applied to quality control. The reason for this may lie in the fact that postponing the arrival time of a product to the metrology tool usually affects the level of quality control. But this is only true under the premise of a low utilization and of a reasonably low waiting time at the metrology tool. As pointed out in Dauzere-Péres et al. (2010) and Rodriguez-Verjan et al. (2013), products that wait too long in the queue for inspection may lose their relevance even under sole quality considerations when new products become available, that are more representative of the current machine status. Closely related is the work of Nemoto et al. (2000) that evaluate the benefits of shorter cycle times to quality due to a shorter feedback loop.

There is ample evidence that variability is one of the main reasons for a production system to perform considerably worse than expected in terms of throughput and flow time (and consequently service level). The behavior of any service or queuing system is determined by the arrival and service processes. In manufacturing systems, it is customary to characterize the station of interest by the two first moments of the inter-arrival time (inter-departure time of the feeding stations) and of the processing time. And so, reducing the line variability is usually achieved by either smoothing the Work-In-Process (WIP) flow, or tackling the process time variability. Because it is in most cases stable, and dictated by technological requirements, the process time on production machines offers a poorer potential in that regard, and most of the managers' efforts naturally target the WIP flow by means of scheduling, dispatching, maintenance scheduling, etc. This is only true for machinery. When human labor is involved, process times may vary widely (especially since breaks are involved). In that regard, note that quality control (metrology) is sometimes one of the last operations in plants that can still be found operated by workers. Because of its tremendous impact on production system performances, numerous researchers have studied ways to model and predict variability (see among others Colledani et al. (2010), Manitz and Tempelmeier (2012), Gershwin (1993), Li and Meerkov (2000), Tan (1999), Assaf et al. (2014) He et al. (2007)). Others like Kalir and Sarin (2009) or Assaf et al. (2014) propose ways to mitigate it, the first paper by coordinating the pace of the machines in the production line, and the second one by using the exceeding capacity on non-bottleneck operations to perform more setups, thus breaking large WIP packets into smaller ones. But again, to the best of our knowledge, metrology has drawn little attention so far in that respect, although there is a major difference between production and metrology operations, i.e. metrology operations are usually not mandatory.

In a former publication (Dauzère-Pérès et al. (2016a)), we formalize a problem for a manufacturing cell where a unique metrology tool controls several heterogeneous production machines. An approach is proposed to determine the sampling periods (the number of products produced between two consecutive inspections) for the different production machines such as to minimize the expected product scrap or rework rate. Production machines are characterized by their failure rate, their throughput rate, and their consumption of the metrology capacity. The resulting problem can be formulated as an optimization problem where the goal is to minimize the expected product loss happening between the machine failure and its detection, subject to the constraint of the metrology capacity, the decision variables being the sampling periods of production machines. The problem was reformulated as a Multiple Choice Knapsack Problem (MCKP), for which we proposed several heuristics based on the work of Sinha and Zoltners (1979) and Pisinger (1995). In Dauzère-Pérès et al. (2016b), we generalize the problem by considering multiple identical metrology tools. In the new problem, decision variables include both the assignment of production machines to metrology tools, and the sampling periods.

In this former body of work, the assumption was made of an immediate answer from the metrology tool, and that production can be synchronized with metrology, and thus that metrology capacity can be fully exploited. More often than not, this is not the case. The variability in the availability of production machines, the fact that production times may differ from one operation to another, and that measurement times are variable make the coordination between arrivals to the metrology operation almost impossible.

Moreover, it is not reasonable to assume that production machines can be stopped or slowed down, and thus production capacity lost, to ensure synchronization of the arrival of products to the metrology tool. As a result, a queue is forming in front of the metrology tool that either impacts production capacity if the production machine waits for the answer from metrology before resuming its activity or impacts quality if, like it is usually the case, production continues in parallel with the product to be measured waiting in front of or being processed on the metrology tool.

In an effort to draw applicable conclusions, in this work, we extend former results to incorporate the quality loss due to the wait and measurement duration at the metrology tool.

In the next section, we formalize the problem at hand. In Section 3, a solution approach is proposed whose performance is then evaluated in Section 4. Finally, in Section 5, we provide managerial insights to evaluate the pertinence of certain level of metrology tool utilization for various industrial scenarios.

2 Mathematical modeling

The general problem and the notations are introduced in Section 2.1, and the contributions in Dauzère-Pérès et al. (2016a) relevant to the present paper are recalled in Section 2.2. Then, our problem with delays in the metrology queue is modeled in Section 2.3.

2.1 Problem description

A group of unreliable production machines send their products to be measured on a unique metrology tool. Each production machine follows its own sampling policy, i.e. each production machine has a sampling period, which is the number of products produced on the machine between two consecutive measures.

The following notations are used in the paper:

- R: Number of production machines,
- TP_r: Throughput rate of production machine r,
- TM_r: Throughput rate of the metrology tool when inspecting products from r,
- p_r: Failure probability (Bernoulli experiment) of production machine r each time it performs a product and if the previous product was good,

- SP_r: Sampling period of production machine r,
- $\lambda_r = \frac{TP_r}{SP_r}$: Inspection rate of machine r, i.e. the rate at which lots are sent to the metrology tool,
- SP^{max}: Upper limit for SP_r over which the risk on production quality is deemed unacceptable by quality managers,
- $\mathcal{S} = \{SP_r; r = 1, \cdots, R\}$: Set of sampling periods.

Note that, if machine r fails, then all following products processed on r are assumed to be defective, i.e. the conditional probability that a product is defective if the previous product is defective is equal to 1. When the product sent to metrology is found defective, the production machine is stopped and repaired after which normal production and inspection cycles resume. Although during repair, no products are sent by the production machine to metrology, we assume these events to be rare enough so that they do not require to be modeled in the metrology utilization.

An alternative approach in practice is to recompute the sampling periods for the remaining machines while one (ore more) machine is being repaired.

We assume the production of a machine in good condition to be perfect, while the production of a defective machine is fully reworked or scrapped.

This assumption is realistic in some industrial settings, such as semiconductor manufacturing, for quality constraints. However, it is possible to extend the analysis in this paper by using a fixed ratio of products that are reworked or scrapped. We also assume there is no difference between the value of products on the different machines. This assumption can be relaxed in our analysis by using a different weight for each production machine associated to the average values of the products performed on the machine. The objective is to minimize the rate at which defective products are produced over all production machines.

The sampling periods SP_r, $\forall\ r = 1, \ldots, R$, are the problem decision variables. They determine both the throughput of bad products from production machine r, and its share in the consumption of the metrology tool capacity, which is denoted by $g_r(SP_r)$ where $g_r(SP_r) = \frac{\lambda_r}{TM_r}$.

2.2 Problem with synchronization of production machines

A sampling period on production machine r is a series of SP_r Bernoulli experiments. A failure occurring during the production of the first product in the sampling period results in SP_r bad products (the number of products until the next inspection takes place). Similarly, a failure occurring during the production of the second product in the sampling period results in $SP_r - 1$ products being reworked or scrapped, and so on. A failure occurring immediately prior an inspection will yield only one bad product. The expected number of bad products from machine r in a sampling period is therefore given by:

$$SP_r p_r + (SP_r - 1)(1 - p_r)p_r + \cdots + 1(1 - p_r)^{SP_r - 1} p_r = p_r \sum_{i=0}^{SP_r - 1} (SP_r - i)(1 - p_r)^i \quad (1)$$

Because λ_r is the arrival rate of products from machine r to the metrology tool, the expected rate of bad products on machine r with sampling period SP_r, called Product Loss, is equal to:

$$PL_r(SP_r) = \lambda_r \cdot p_r \cdot \sum_{i=0}^{SP_r-1} (SP_r - i)(1-p_r)^i \tag{2}$$

Note that, in (2), PL_r only depends on the sampling period SP_r of machine r, and thus can be computed a priori for each possible value of SP_r. In Section 2.3, when delays in the metrology queue are considered, we show that the product loss related to machine r depends on the sampling periods of all machines, i.e. of $\mathcal{S}$.

It is possible to formalize the problem as the Integer Linear Program (ILP) below by introducing a binary variable u_r^s which is equal to 1 if the selected sampling period for machine r is equal to s, and 0 otherwise.

$$\min \sum_{r=1}^{R} \sum_{s=1}^{SP^{max}} PL_r(s) u_r^s \tag{3}$$

s.t.

$$\sum_{r=1}^{R} \sum_{s=1}^{SP^{max}} g_r(s) u_r^s \leq 1 \tag{4}$$

$$\sum_{s=1}^{SP^{max}} u_r^s = 1, \quad r = 1, \ldots, R \tag{5}$$

$$u_r^s \in \{0,1\}, \qquad r = 1, \ldots, R; \; s = 1, \ldots, SP^{max} \tag{6}$$

Constraint (4) ensures that the metrology capacity is satisfied, and Constraints (5) that one and only one sampling period s is selected for each machine r.

To solve the ILP above efficiently, Heuristic $H_{2/3}$ in Dauzère-Pérès et al. (2016a) is proposed, which will be used within the solution approach proposed in Section 3.

2.3 Problem with delays in metrology queue

Differently than in Dauzère-Pérès et al. (2016a), we now consider that a queue is forming at the metrology tool that delays its response (see Figure 1). As defined earlier, the arrival rate of products from machine r to the metrology tool is λ_r, and so the total arrival rate to the metrology tool is:

$$\lambda(\mathcal{S}) \triangleq \sum_{r=1}^{R} \lambda_r$$

Consequently, the expected proportion of the production waiting to be checked at the metrology tool coming from machine r is $\lambda_r/\lambda(\mathcal{S})$. The expected service rate at the metrology tool is the weighted harmonic mean of rates TM_r for the different production

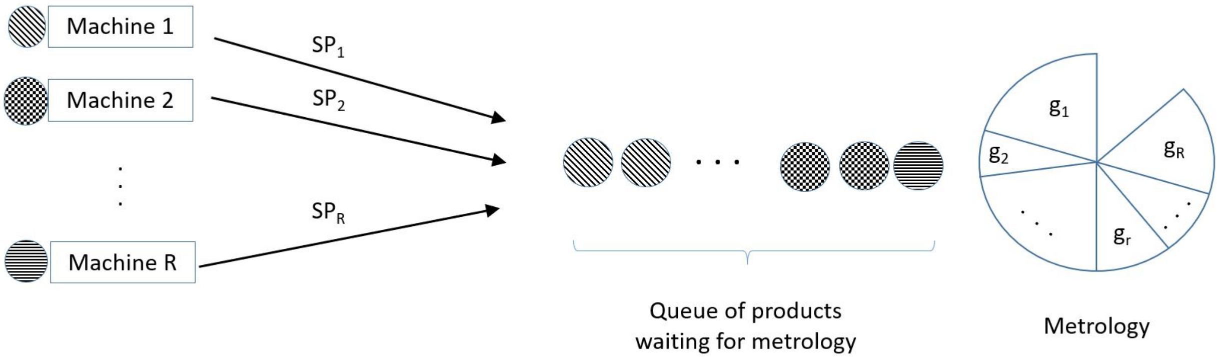

Figure 1: Sampling periods and metrology capacity

machines:

$$\mu(\mathcal{S}) \triangleq \frac{\displaystyle\sum_{r=1}^{R} \lambda_r}{\displaystyle\sum_{r=1}^{R} \frac{\lambda_r}{TM_r}} \tag{7}$$

Let us assume that, due to the lack of synchronization between production machines but also to the workshop inherent variability, both the inter-arrival time to the metrology tool and the service time on the metrology tool are sampled from general distributions. With the expected traffic intensity, which measures the congestion of the system, defined as

$$\rho(\mathcal{S}) \triangleq \frac{\lambda(\mathcal{S})}{\mu(\mathcal{S})} = \sum_{r=1}^{R} \frac{\lambda_r}{TM_r},$$

the expected sojourn time of lots sent to metrology is given by Kingman's approximation (see Kingman (1962)) for G/G/1 queues:

$$W(\mathcal{S}) \approx \frac{1}{\mu(\mathcal{S})} \left[\left(\frac{\rho(\mathcal{S})}{1-\rho(\mathcal{S})} \right) \left(\frac{c_a^2 + c_s^2}{2} \right) + 1 \right] \tag{8}$$

where c_a and c_s are the coefficients of variation of the inter-arrival time and of the service time, respectively.

The number of products produced on machine r between two inspections is equal to SP_r when there is no machine failure, i.e. with probability $(1-p_r)^{SP_r}$. However, differently from the analysis in Section 2.2, a failure occurring when processing any of the SP_r products also results in a number of bad products $\lceil W(\mathcal{S}) \cdot TP_r \rceil$ produced during the time $W(\mathcal{S})$ spent in the queue by the product sent to be measured.

An inspection period begins with machine r performing properly (either from a negative inspection or the end of a repair) and ends either with the next inspection or by a machine stoppage. The inspection period duration is SP_r/TP_r when production is good, i.e. with probability $(1-p_r)^{SP_r}$. The inspection period duration is $(SP_r + \lceil W(\mathcal{S}) \cdot TP_r \rceil)/TP_r$ when a failure occurs, i.e. with probability $1-(1-p_r)^{SP_r}$. The expected inspection period duration is therefore:

$$
\begin{aligned}
&(1-p_r)^{SP_r} \cdot \frac{SP_r}{TP_r} + (1-(1-p_r)^{SP_r}) \cdot \frac{SP_r + \lceil W(\mathcal{S}) \cdot TP_r \rceil}{TP_r} \\
&= \frac{SP_r + (1-(1-p_r)^{SP_r}) \cdot \lceil W(\mathcal{S}) \cdot TP_r \rceil}{TP_r} \qquad (9)
\end{aligned}
$$

And the average arrival rate of production from machine r to the metrology until r is stopped because of a failure is given by:

$$
\frac{TP_r}{SP_r + (1-(1-p_r)^{SP_r}) \cdot \lceil W(\mathcal{S}) \cdot TP_r \rceil} \qquad (10)
$$

Following which, the overall expected rate of defective products, i.e. the product loss, of machine r under an inspection policy using the set of sampling periods $\mathcal{S} = \{SP_1, \ldots, SP_R\}$, is the average arrival rate of production from machine r until stoppage (10) multiplied by the sum of the expected product loss between two inspections (1) and the expected product loss during the inspection sojourn time $(1-(1-p_r)^{SP_r}) \cdot \lceil W(\mathcal{S}) \cdot TP_r \rceil$, i.e.:

$$
PL_r(\mathcal{S}) = \frac{TP_r \cdot \left[p_r \cdot \sum_{i=0}^{SP_r-1} (SP_r - i)(1-p_r)^i + (1-(1-p_r)^{SP_r}) \cdot \lceil W(\mathcal{S}) \cdot TP_r \rceil \right]}{SP_r + (1-(1-p_r)^{SP_r}) \cdot \lceil W(\mathcal{S}) \cdot TP_r \rceil} \qquad (11)
$$

The total expected product loss from a set of sampling periods $\mathcal{S}$ is given by:

$$
PL(\mathcal{S}) = \sum_{r=1}^{R} PL_r(\mathcal{S}) \qquad (12)
$$

Note that, because PL_r depends on the set of all sampling periods and not only on the sampling period of machine r, $PL(\mathcal{S})$ is not separable by production machine.

The optimization problem (P) can be written as follows:

$$
\begin{aligned}
&\min \; PL(\mathcal{S}) \\
&\text{s.t.} \\
&\sum_{r=1}^{R} g_r(SP_r) \; < \; 1 \\
&SP_r \in \{1, \ldots, SP^{max}\}, \quad r = 1, \ldots, R
\end{aligned}
$$

Note that without variability, i.e. if $c_a^2 + c_s^2 = 0$, then $W(\mathcal{S}) = \frac{1}{\mu(\mathcal{S})}$ with (8). However, in this special, extreme case, using (8) is no longer relevant since products of different machines are no longer interconnected through a common queue, making an approximation for the sojourn time unnecessary. Hence, the sojourn time of each product is separable and is equal to $\frac{1}{TM_r}$, which should replace $W(\mathcal{S})$ in (11). Even in this case, (12) is larger than (2).

As already mentioned, due to the loss associated with products produced during the additional sojourn time in the inspection queue of products to be measured, the target function in the optimization problem (P) is not separable per production machine, and Heuristic $H_{2/3}$ in Dauzère-Pérès et al. (2016a) is not directly applicable, i.e. it is not possible to write the problem as the ILP presented in Section 2.2. However, in the next section, we propose a solution approach based on Heuristic $H_{2/3}$ to solve (P).

3 Solution approach

In a variable environment where no synchronization is possible between the metrology availability and the arrival of products to be measured, the expression of $PL(\mathcal{S})$ shows that there is a tradeoff between the benefits associated with decreasing the sampling periods of the different production machines and the product loss due to the resulting waiting at the metrology tool. As an example of such a tradeoff, Figure 2 presents, for a given set of parameters to be introduced in Section 4, the evolution of the optimal value of the product loss $PL(\mathcal{S})$ as a function of the allowed metrology utilization. For each level of metrology utilization, a set of optimal sampling periods was determined, and the resulting product loss calculated. Using the metrology tool to its full capacity leads to an infinite queue, and thus to losing all products on a machine when there is a failure. Reducing the metrology tool utilization by increasing the sampling periods helps to get faster answers from the metrology tool, but potentially at the expense of a larger product loss because less products are inspected. This tradeoff is an integral part of the cost of quality of processes, that includes prevention costs, appraisal costs and internal and external failure costs.

Let us simplify the problem by hierarchically considering the questions of the right level of utilization for the metrology tool and of the sampling period for each production machine. Our approach iteratively uses the solution of a modified version of the Integer Linear Program (ILP) recalled in Section 2.2 and solved in Dauzère-Pérès et al. (2016a). This ILP, denoted IP(c, W), where c is the capacity limitation and W is a given queue sojourn time, is written below.

$$\min \sum_{r=1}^{R} \sum_{s=1}^{SP^{max}} PL_r(s, W) u_r^s$$

s.t.

$$\sum_{r=1}^{R} \sum_{s=1}^{SP^{max}} g_r(s) u_r^s \leq c$$

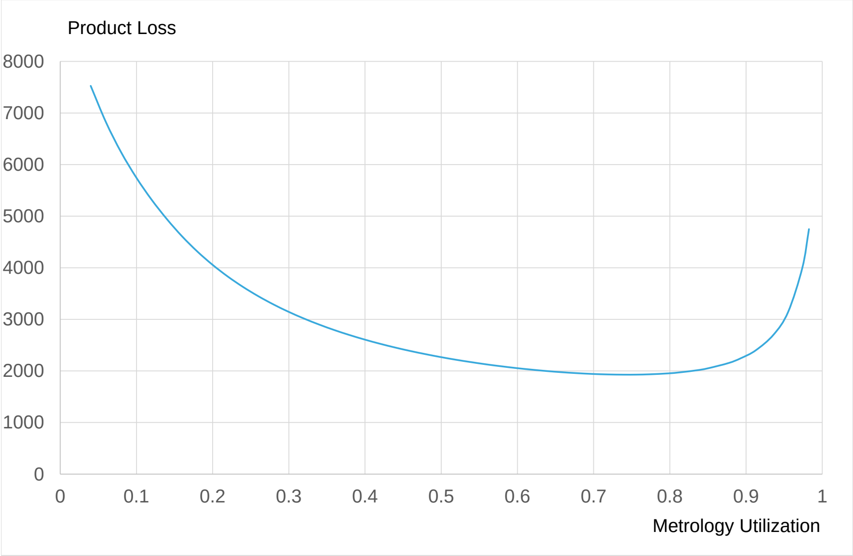

Figure 2: Product Loss as a function of metrology utilization, $R = 10$, $p_{max} = 0.05$, $TP_{min} = 900$, $RTP/TM_r = 10$, $v = 0.8$.

$$\sum_{s=1}^{SP^{max}} u_r^s = 1, \quad r = 1, \ldots, R$$

$$u_r^s \in \{0, 1\}, \qquad r = 1, \ldots, R; \; s = 1, \ldots, SP^{max}$$

where

$$PL_r(s, W) = \frac{TP_r \cdot \left[p_r \cdot \sum_{i=0}^{s-1} (s-i)(1-p_r)^i + (1-(1-p_r)^s) \cdot \lceil W \cdot TP_r \rceil \right]}{s + (1-(1-p_r)^s) \cdot \lceil W \cdot TP_r \rceil}. \tag{13}$$

$PL_r(s, W)$ in (13) corresponds to $PL_r(\mathcal{S})$ in (11), where $W(\mathcal{S})$ is fixed to W and s is chosen as the sampling rate SP_r of machine r. Note that $SP_r = s$ if $u_r^s = 1$ in IP(c, W).

The differences between the problem in Dauzère-Pérès et al. (2016a) and IP(c, W) are the capacity (1 vs. c) and (13) which determines the product loss for production machine r under an assumption of delays in metrology independent from $\mathcal{S}$.

The objective function in IP(c, W) is separable when W is given, since $PL_r(s, W)$ can be precomputed for each s and each r, which allows a good solution to be found in a competitive time using Heuristic $H_{2/3}$ in Dauzère-Pérès et al. (2016a). The Product Balancing Heuristic (PLB) presented in Algorithm 1 iteratively solves IP(c, W), starting with a capacity of 1, until $PL(\mathcal{S})$ no longer decreases. In each iteration k, a first set $\mathcal{S}^k$ of sampling periods is found in Step 4 assuming no waiting time at the metrology tool, i.e. $W = 0$. The algorithm then iteratively looks for better values of $\mathcal{S}^k$, by updating W in Step 11 each time a new set $\mathcal{S}$ is determined in Step 10, and up to ten iterations. In our experiments, much fewer iterations are usually needed for the product loss to converge or to cycle. This is why Steps 15 and 16 are introduced. The product loss is computed in Step 12 and the best set of sampling periods $\mathcal{S}^k$ at iteration k is updated in Step 14. Finally, the capacity is set in Step 19 to a value slightly smaller than the utilization resulting from $\mathcal{S}^k$, in order to trigger a new solution in the next iteration.

Algorithm 1 Product Loss Balancing (PLB) Heuristic

```
 1: Set $c^0 = 1$, $k = 0$ (iteration number), and $\epsilon$ to a very small value
 2: repeat
 3:    $k \leftarrow k + 1$
 4:    Determine $\mathcal{S}^k$ by solving IP($c^k, 0$) using $H_{2/3}$ from Dauzère-Pérès et al. (2016a)
 5:    $W = W(\mathcal{S}^k)$ (using (8))
 6:    Calculate $PL(\mathcal{S}^k)$ (using (12))
 7:    $i = 0$
 8:    repeat
 9:       $i \leftarrow i + 1$
10:       Determine $\mathcal{S}$ by solving IP($c^k, W$) using $H_{2/3}$ from Dauzère-Pérès et al. (2016a)
11:       $W = W(\mathcal{S})$ (using (8))
12:       Calculate $PL(\mathcal{S})$ (using (12))
13:       if $PL(\mathcal{S}) < PL(\mathcal{S}^k)$ then
14:          $\mathcal{S}^k \leftarrow \mathcal{S}$
15:       else if $PL(\mathcal{S}) = PL(\mathcal{S}^k)$ then
16:          $i = 10$
17:       end if
18:    until $i = 10$
19:    $c^{k+1} = \sum_{r=1}^{R} g(SP_r^k) - \epsilon$ where $\mathcal{S}^k = \{SP_1^k, \dots, SP_R^k\}$
20: until $PL(\mathcal{S}^k) > PL(\mathcal{S}^{k-1})$
21: Return $\mathcal{S}^{k-1}$, $\sum_{r=1}^{R} g(SP_r^{k-1})$, and $PL(\mathcal{S}^{k-1})$
```

The value of ϵ is chosen small enough so that, if there is more than one solution between the solutions obtained with c^{k-1} and c^k, it does not offer any meaningful difference with either of them. In the numerical experiments of the following section, $\epsilon = 0.001$.

4 Computational analysis

Section 4.1 presents how the experiments were designed. Section 4.2 shows how a simulation was conducted to study the variability related to the lack of synchronization between production machines in a deterministic setting. Finally, the numerical results are presented and discussed in Section 4.3.

4.1 Structure of the experiments

The behavior of the algorithm presented in Section 3 is studied using a rather large set of scenarios to generate test instances that correspond to replications of the scenarios. Each scenario is defined by a given range for each of the problem parameters, and 10 instances are created for each scenario by randomly generating the parameters within the specified ranges. The number of production machines R is chosen in the set $\{10, 20, 40\}$. We did not consider scenarios with too few machines because the number of production machines is in practice often quite large (here are some examples in a semiconductor manufacturing facility: 1 metrology tool for 10 production machines, 3 metrology tools for about 80 production machines and 2 metrology tools for about 40 production machines). The parameters for each of the production machines are determined as follows. The failure probabilities p_r are generated from a uniform distribution $U[p_{min}, p_{max}]$, where p_{max} is chosen in the set $\{0.05, 0.2\}$ and p_{min} is kept constant ($p_{min} = 0.01$). The throughput rate TP_r is generated from a distribution $U[TP_{min}, TP_{max}]$, where $TP_{max} = 1,000$ and TP_{min} is chosen in the set $\{100, 900\}$. The measurement rate TM_r is determined using the ratio $\frac{R \cdot \overline{TP}}{TM_r}$ chosen from the set $\{5, 10, 30\}$, where $\overline{TP}$ is the average throughput rate for the considered scenario.

The variability factor is more complex to handle in the framework of this experimentation since it is affected by the choice of sampling periods, and by the problem parameters. Let us denote by v^2 the variability factor in Kingman's approximation (Kingman (1962)) $\left(v^2 = \frac{c_a^2+c_s^2}{2}\right)$. In Hopp and Spearman (2011), the coefficient of variation (and incidentally v) is classified as "Low Variability" if it is smaller than 0.75, "Medium Variability" if it is between 0.75 and 1.33, and "High Variability" if it is larger than 1.33.

A priori, we wanted to cover a wide range of variability factors, although some of them are not realistic. To overcome this problem, v was first chosen from the set {0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 2}. As seen above, for each scenario, 10 instances were randomly generated, each with different values of p_r and TP_r. This led to 504 scenarios characterized by different sets of parameters, and a total of 5,040 different instances. We performed the full set of experiments using heuristic PLB to determine the set of sampling periods for each instance, knowing that some scenarios are characterized by a variability factor that is too small to be feasible. We then studied the feasibility of each scenario based on the resulting sojourn time computed using (8) and on the result of a simulation described in the following section.

4.2 Simulation of sojourn times in the deterministic case

For each of the 5,040 instances, we estimated a lower bound of the sojourn time by simulating a deterministic arrival process to the metrology tool from the relevant number of production machines using the set of sampling periods $\mathcal{S}$ determined by heuristic PLB, followed by a deterministic service (measurement) process, both with the scenario characteristics. The simulation is performed until 100,000 products have been sent to the metrology tool from one production machine.

This deterministic simulation provides an average sojourn time which represents a minimum in terms of variability. Because, in a deterministic framework, the synchronization between departures from production machines impacts the results, 50 replications were generated, each with different randomly generated machine starting points in time, and the sojourn time was averaged over these replications. An instance for which the estimated sojourn time resulting from (8) is lower than the average sojourn time of the simulation is deemed unrealistic. For our parameters, the results showed a very low sensitivity to the synchronization (the results are very similar between the 50 simulations for a given instance), but showed some dependence on the scenario characteristics, mainly on the number of machines. Table 1 shows, for each combination of number of machines and variability factor, the proportion of cases in which the sojourn time estimated using (8) is larger than the sojourn time estimated with the simulation, in both cases with the set of sampling periods $\mathcal{S}$ determined using heuristic PLB.

Table 1: Lower bound for variability factor v as a function of the number of production machines R

	R		
v	10	20	40
0	0%	0%	0%
0.05	0 %	0 %	0 %
0.1	0 %	0 %	0 %
0.2	0 %	0 %	0 %
0.3	0 %	0 %	0 %
0.4	36.7 %	0 %	0 %
0.5	99.2 %	95.8 %	80 %
0.6	99.2 %	96.7 %	92.5 %
0.8	100 %	99.2 %	95 %
1	100 %	100 %	95.8 %
1.2	100 %	100 %	100 %
1.4	100 %	100 %	100 %
1.6	100 %	100 %	100 %
2	100 %	100 %	100 %

Adopting a cutoff point of 95%, we conclude that only scenarios with a coefficient of variability $v \geq 0.5$ for 10 and 20 machines, and with $v \geq 0.8$ for scenarios with 40 machines, i.e. 2,640 instances in total, are considered realistic. The remainder of our analysis is based on these instances exclusively.

4.3 Numerical results

To first address the behavior of the solution provided by our approach, each instance is associated to the median of the metrology capacity utilization and product loss in the first solution, which strives to fill metrology capacity and disregards any delay in the result, and in the solution provided by the PLB heuristic. Because the product loss rises sometimes steeply as metrology utilization gets closer to 1, small differences in metrology utilization sometimes lead to large differences in product loss, which is why we preferred to use the median of the performance values of interest instead of their averages. We then calculated the ratio of utilization (utilization optimized using PLB vs. initial utilization with metrology capacity set to 1) and the resulting reduction in product loss (ratio of $PL(\mathcal{S})$ optimized using PLB vs. initial $PL(\mathcal{S})$ with metrology capacity set to 1). In theory, filling the metrology capacity should result in an infinite queue length, and therefore a product loss equal to $\sum_{r=1}^{R} TP_r$. However, in our case, because the metrology utilization is consumed in discrete quanta, the optimization results in a variety of product loss values depending on how close the solution is to full metrology utilization, and on the variability factor. Figure 3 plots the product loss reduction obtained by the PLB heuristic versus the metrology utilization ratio. We can observe a very broad range of responses from the PLB heuristic.

In some instances, the reduction in product loss is limited (as low as 10%) while the metrology capacity stays highly utilized (90% of the original utilization). In other extreme cases, most of the product loss is created by the delay in metrology and reducing the metrology utilization allows for a reduction of most of the product loss (up to about 80% of the original product loss). Note also that the smallest utilization ratios (larger reduction in utilization) are just above 53%, which is quite low in regular industrial settings, and that there is no case where fully utilizing the metrology seems to make sense.

The way a specific scenario responds to a reduction in metrology utilization is related to its characteristics in general, but we know for a fact that it depends on the variability. Figure 4 plots the metrology utilization ratio as a function of the variability factor, which prompts two main observations. First, the ratio decreases, which means that, unsurprisingly, the metrology capacity utilization should be lower as the variability increases. Second, note that the range of metrology utilization ratios for a specific variability factor is large enough to conclude that variability alone cannot explain a specific result.

Having established how our approach behaves in general, in the next section, we propose to draw applicable practical guidelines from our computational results.

5 Managerial Recommendations

Is there enough information in the characteristics of a set of production machines controlled by a single metrology tool to correctly set its capacity utilization at an adequate level in order to minimize the product loss? This is the question we address in this section based on our experimental results. To do so, we defined and used a prediction model, namely a

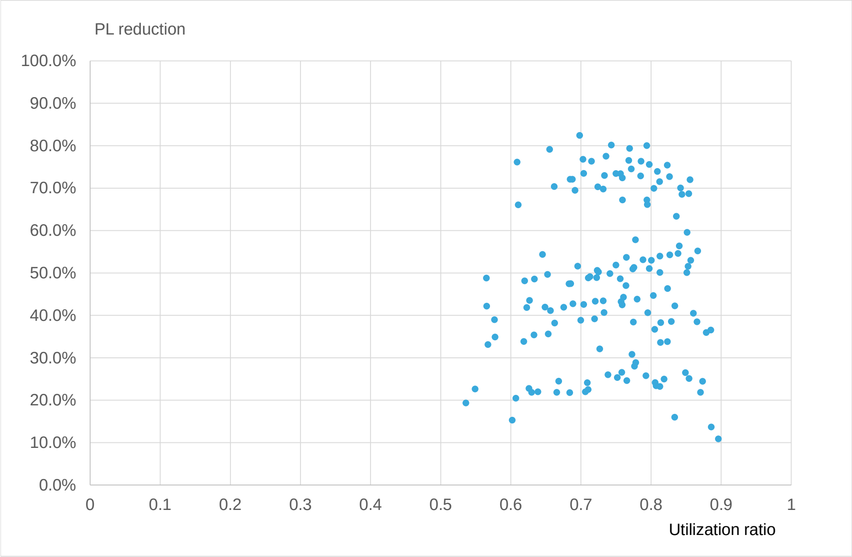

Figure 3: Product loss reduction as a function of metrology utilization reduction

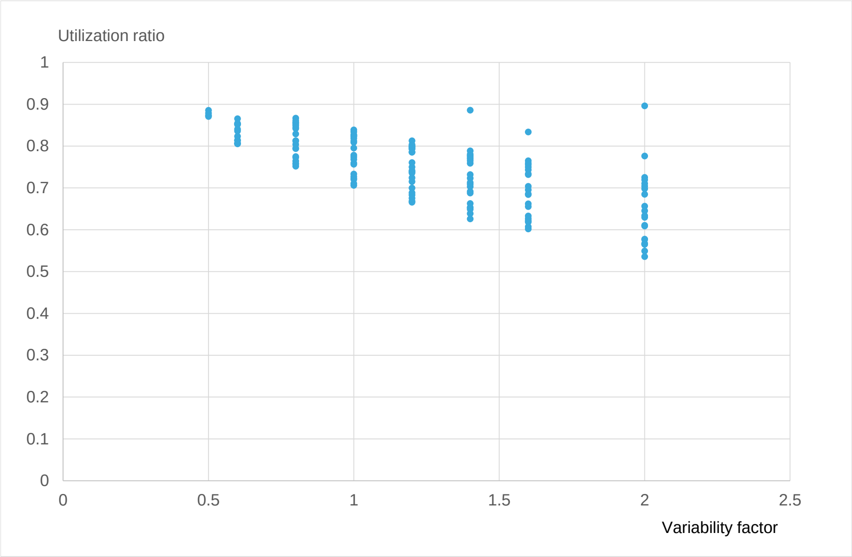

Figure 4: Metrology utilization reduction as a function of variability

regression tree, for the metrology utilization. The following analysis was implemented in R 3.4.3. The scenario characteristics, namely the machine reliability (p_{max}), production rate (TP_{min}), measurement rate relative to the production rate ($\frac{R \cdot \overline{TP}}{TM_r}$), and the variability factor (v) were defined as potential predictors. The data are composed of 2,640 tuples, one for each instance. Each tuple is composed by the instance parameters and the metrology utilization giving the best solution (smallest PL) with PLB.

To properly build and test a regression tree, the following procedure was performed. The tuples were first randomly shuffled, then randomly segregated into a training set (70% of the tuples) and a test set (the remaining 30% of tuples). A regression tree model was trained on the training set. Then, the quality of the model prediction was evaluated on the test set, based on the explained variance.

We applied this procedure 20 times, among which the average and minimum of the explained variance were 90.1% and 87.3%, respectively. In all trees, the only predictors picked by the algorithm are the coefficient of variation and the number of machines. The split nodes reveal that the variability factor and the number of machines are always the two parameters yielding the largest improvement, systematically about one order of magnitude above any other potential predictor.

The relative contribution of the two predictors is not equivalent and the coefficient of variability is systematically twice as important as the number of machines (67% vs. 33% respectively). It also appears that the resulting trees are up to four levels deep and share the same basic structure (splitting conditions) of their upper levels. Only the final splits are sometimes slightly different. This allows us to postulate that, given the number of machines served by the metrology tool and the variability factor of such a system, it should be possible to set the level of utilization of the metrology tool to minimize the risk.

Because the model proved mostly insensitive to the random splits of the training and test sets, and in order to derive simple recommendations for practitioners, we trained the regression tree model with the totality of the data, requiring a maximum depth of 3. The model itself is presented in Figure 5. In each leaf, the main characteristics of the sub-population it represents are shown, namely the average optimal metrology utilization and its size relative to all the tuples. Because of the limitation on the tree depth, the explained variance is slightly lower at 85.2%.

These results allow us to provide simple recommendations to managers with a reasonable level of confidence. It is especially interesting that the machine reliability, which is typically difficult to assess in most situations, does not appear to contribute to the characterization of optimal solutions.

The number of production machines feeding the metrology tool is an obvious metric. The managing team is therefore left with the task of evaluating the variability coefficient by measuring the coefficient of variation of the inter-arrival time to the metrology tool, and of the measurement times. The regression tree shows the average value for the utilization of each leaf, but without range of values in the leaf population. This alone does not allow to draw applicable guidelines. In order to propose meaningful ranges of utilization, we built 95% confidence intervals in each leaf. Without assuming any prior distribution for

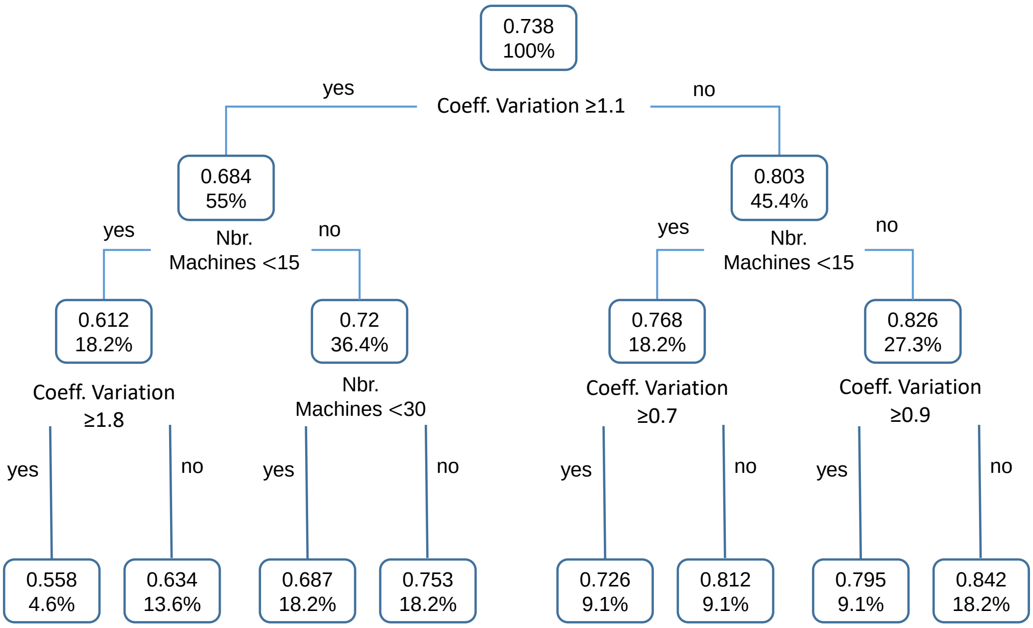

Figure 5: Prediction tree for metrology utilization

the results, the 2.5% and 97.5% percentiles were calculated for the utilization in each leaf. We are now able to present guidelines for managers in Table 2. Each leaf in the tree is presented as a condition rule on the two sole predictors (variability factor and number of machines), and the recommended range of metrology utilization, corresponding to the aforementioned percentiles, is provided.

Table 2: Recommended metrology capacity utilization

Rule	1	2	3	4	5	6	7	8
Variability factor	> 1.8	$1.1 - 1.8$	> 1.1	> 1.1	$0.7 - 1.1$	< 0.7	$0.9 - 1.1$	< 0.9
Number of machines	< 15	< 15	$15 - 30$	> 30	< 15	< 15	> 15	> 15
Utilization range	51% – 58%	58% – 69%	60% – 75%	68% – 81%	68% – 77%	77% – 84%	74% – 84%	79% – 88%

Table 2 prompts a few observations. Unsurprisingly, for a given number of machines, the recommended metrology capacity utilization decreases with the variability factor. For example, rules 2 and 5 which share the same range of number of machines (< 15) show how a higher variability factor (between 1.1 and 1.8 for rule 2) results in a lower recommended metrology utilization. However, it is important to keep in mind that an inbound flow of products with low variability normally requires all production machines to produce at a similar rate with little variability. Also, the measurement times themselves must be similar enough and the metrology tool needs to be reliable.

The recommended capacity utilization also increases with the number of production machines. The analysis of sojourn times presented in Section 4.2 shows how the number of production machines impacts the variability of the inter-arrival time to the metrology tool (recall that the inter-departure time from each production machine r is deterministic). But the increase of the number of production machines cannot be the reason for the increase of the recommended capacity utilization, since the variability is fully contained in parameter v in our formulation. Another possibility is that, with more production machines requiring service from the metrology tool, our algorithm naturally assigns larger sampling periods for which the metrology capacity is consumed in smaller ratios, making it easier for the algorithm to reach a better solution to problem (P). However, the levels of stress on the metrology tool are maintained comparable for different number of machines by using the ratio $\frac{R \cdot TP}{TM_r}$ to determine TM_r, and the sampling periods resulting from our experiments are indeed very close, with averages of 33.8, 32.7 and 34.1 for 10, 20 and 40 machines, respectively. We found that the reason for the positive impact of a larger number of machines on the system lies in the added flexibility offered to the algorithm, which allows for a larger range of possible assignments of sampling periods SP_r for each scenario. Indeed, while on the average the values of the sampling periods SP_r are similar for different number of machines, the standard deviation of the sampling periods for each scenario is fairly different, with average values of 23.2, 29.4 and 32.9 for 10, 20 and 40 machines, respectively.

For the most extreme combinations of R and v, the recommended metrology capacity utilization drops significantly, with extreme values between 51% and 58%.

This may come as a surprise as low levels of capacity utilization, in particular for human operators, are usually difficult to accept. This is not to say that a metrology tool that is idle about half of the time is to be systematically ignored, but managers may well need to acknowledge that, in some cases, this is actually the operational mode that minimizes loss on production quality.

We propose to use these recommendations as follows in industrial settings:

- Measure the production machine characteristics, failure rates, production rates, etc.
- Observe the combined interarrival process to the metrology tool, taking into accounts all disturbances, and calculate its coefficient of variation c_a,
- Observe the mixed service time at the metrology tool taking into accounts all disturbances, and calculate its coefficient of variation c_s,
- Calculate the overall variability factor $v = \sqrt{\frac{c_a^2+c_s^2}{2}}$,
- Estimate the metrology tool capacity utilization,
- Check if, based on the number of production machines R and the variability factor v, the metrology tool capacity utilization in use is reasonable according to Table 2,
- Scrutinize the workshop and consider adjusting the metrology tool capacity utilization if it looks inadequate.

Let us recall the industrial case presented in Dauzère-Pérès et al. (2016a), characterized by 34 production machines. In this example, the recommendations would be the following: If the variability factor v is lower than 0.9, then a metrology tool capacity utilization between 79% and 88% can be safely reached. Any figure over that range could mean an increased quality loss due to the products waiting in the metrology queue. Below it, the metrology tool is probably under-utilized, and tighter quality control could be achieved. For $0.9 < v < 1.1$, a lower level of utilization is advised $(74\% - 84\%)$ while, for $v > 1.1$ the right metrology tool capacity utilization should be between 68% and 81%.

Finally, note that although we are able to recommend certain levels of metrology tool utilization, this alone is in no way enough to ensure an optimal operation of the inspection process. The metrology utilization is a direct result of the sampling periods, which means that a seemingly adequate level of utilization can hide a set of inadequate sampling rates, resulting in a product loss that is not optimal. The opposite, however, is not true, and here lies the interest of our recommendations: What Table 2 does provide is a quick and simple way for the practitioner to determine if a metrology tool has a capacity utilization that is "wrong" and would impact production quality, regardless of the sampling periods. In order to actually reach an optimal operational mode, a careful determination of the sampling periods is still required, which can be obtained by using Heuristic PLB proposed in Section 3.

6 Conclusion

In this paper, we studied the impact of variability when managing metrology capacity with the goal of proposing guidelines to managers. After modeling how variability is increasing the loss on production yield, an approach is presented to optimize sampling periods and metrology capacity utilization so as to minimize product loss. This approach uses the best algorithm proposed in Dauzère-Pérès et al. (2016a) for a given metrology capacity. Extensive computational experiments were conducted to analyze the impact of various key characteristics of the problem, and a prediction model was used to analyze our computational results and define which parameters are the most important. We were then able to derive managerial recommendations for metrology capacity utilization depending on two parameters: The variability factor and the number of production machines.

We hope that the insights in this paper will help managers to better understand the interest of formalizing the impact of variability on metrology capacity. Metrology being a mandatory operation in modern high-tech manufacturing systems, defining the right metrology capacity to assign to production machines is thus an important problem. We also believe that Table 2 should be helpful for practitioners to define the right metrology capacity utilization to use in approaches optimizing sampling periods such as the ones in Bettayeb et al. (2012), Nduhura-Munga et al. (2015), Rodriguez-Verjan et al. (2015) and Dauzère-Pérès et al. (2016a). This should lead to the definition of more complex optimization models where the metrology capacity becomes a decision variable. Solving the resulting models will probably require dedicated solution approaches to be developed.

Acknowledgments

This research was partly carried out whilst the second author was visiting the Department of Manufacturing Sciences and Logistics of the Center of Microelectronics in Provence of the Ecole des Mines de Saint-Etienne, Gardanne, France in October 2017 and September 2018.

References

Assaf, R., Colledani, M., Matta, A., 2014. Analytical evaluation of the output variability in production systems with general markovian structure. OR spectrum 36 (3), 799–835.

Bettayeb, B., Bassetto, S., Vialletelle, P., Tollenaere, M., 2012. Quality and exposure control in semiconductor manufacturing. part I: Modelling. International Journal of Production Research 50 (23), 6835–6851.

Bouslah, B., Gharbi, A., Pellerin, R., 2016. Integrated production, sampling quality control and maintenance of deteriorating production systems with AOQL constraint. Omega 61, 110–126.

Colledani, M., Matta, A., Tolio, T., 2010. Analysis of the production variability in multistage manufacturing systems. CIRP Annals-Manufacturing Technology 59 (1), 449–452.

Colledani, M., Tolio, T., 2011. Integrated analysis of quality and production logistics performance in manufacturing lines. International Journal of Production Research 49 (2), 485–518.

Dauzère-Pérès, S., Hassoun, M., Sendon, A., 2016a. Allocating metrology capacity to multiple heterogeneous machines. International Journal of Production Research 54 (20), 6082–6091.

Dauzère-Pérès, S., Hassoun, M., Sendon, A., 2016b. Optimizing capacity assignment of multiple identical metrology tools. In: Winter Simulation Conference 2016 (WSC 2016). IEEE, pp. 2709–2718.

Dauzere-Péres, S., Rouveyrol, J.-L., Yugma, C., Vialletelle, P., 2010. A smart sampling algorithm to minimize risk dynamically. In: 2010 IEEE/SEMI Advanced semiconductor manufacturing conference (ASMC 2010). IEEE, pp. 307–310.

Gershwin, S., 1993. Variance of output of a tandem production system. Queueing Networks with Finite Capacity, 291–304.

Gilenson, M., Hassoun, M., Yedidsion, L., 2015. Setting defect charts control limits to balance cycle time and yield for a tandem production line. Computers & Operations Research 53, 301–308.

Goyal, S., Gunasekaran, A., Martikainen, T., Yli-Olli, P., 1993. Integrating production and quality control policies: A survey. European journal of operational research 69 (1), 1–13.

He, X.-F., Wu, S., Li, Q.-L., 2007. Production variability of production lines. International Journal of Production Economics 107 (1), 78–87.

Hendry, L., Kingsman, B., Cheung, P., 1998. The effect of workload control (WLC) on performance in make-to-order companies. Journal of Operations Management 16 (1), 63–75.

Hopp, W. J., Spearman, M. L., 2011. Factory physics. Waveland Press.

Kalir, A. A., Sarin, S. C., 2009. A method for reducing inter-departure time variability in serial production lines. International Journal of Production Economics 120 (2), 340–347.

Kingman, J., 1962. On queues in heavy traffic. Journal of the Royal Statistical Society. Series B (Methodological), 383–392.

Kingsman, B., Hendry, L., 2002. The relative contributions of input and output controls on the performance of a workload control system in make-to-order companies. Production Planning & Control 13 (7), 579–590.

Lee, S., Lee, T., Liao, J., Chang, Y., 2003. A capacity-dependence dynamic sampling strategy. In: IEEE international symposium on semiconductor manufacturing. pp. 312–314.

Li, J., Meerkov, S. M., 2000. Production variability in manufacturing systems: Bernoulli reliability case. Annals of Operations Research 93 (1-4), 299–324.

Manitz, M., Tempelmeier, H., 2012. The variance of inter-departure times of the output of an assembly line with finite buffers, converging flow of material, and general service times. OR spectrum 34 (1), 273–291.

Nduhura-Munga, J., Dauzère-Pérès, S., Vialletelle, P., Yugma, C., 2012. Industrial implementation of a dynamic sampling algorithm in semiconductor manufacturing: Approach and challenges. In: Winter Simulation Conference 2012 (WSC 2012). IEEE, pp. 1–9.

Nduhura-Munga, J., Dauzère-Pérès, S., Yugma, C., Vialletelle, P., 2015. A mathematical programming approach for optimizing control plans in semiconductor manufacturing. International Journal of Production Economics 160, 213–219.

Nduhura-Munga, J., Rodriguez-Verjan, G., Dauzère-Pérès, S., Yugma, C., Vialletelle, P., Pinaton, J., 2013. A literature review on sampling techniques in semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 26 (2), 188–195.

Nemoto, K., Akcali, E., Uzsoy, R. M., 2000. Quantifying the benefits of cycle time reduction in semiconductor wafer fabrication. IEEE Transactions on Electronics Packaging Manufacturing 23 (1), 39–47.

Orcun, S., Uzsoy, R., Kempf, K. G., 2009. An integrated production planning model with load-dependent lead-times and safety stocks. Computers & Chemical Engineering 33 (12), 2159–2163.

Pisinger, D., 1995. A minimal algorithm for the multiple-choice knapsack problem. European Journal of Operational Research 83 (2), 394–410.

Rodriguez-Verjan, G. L., Dauzère-Pérès, S., Housseman, S., Pinaton, J., 2013. Skipping algorithms for defect inspection using a dynamic control strategy in semiconductor manufacturing. In: Winter Simulation Conference 2013 (WSC 2013). IEEE, pp. 3684–3695.

Rodriguez-Verjan, G. L., Dauzère-Pérès, S., Pinaton, J., 2015. Optimized allocation of defect inspection capacity with a dynamic sampling strategy. Computers & Operations Research 53, 319–327.

Shanoun, M., Bassetto, S., Bastoini, S., Vialletelle, P., 2011. Optimisation of the process control in a semiconductor company: model and case study of defectivity sampling. International Journal of Production Research 49 (13), 3873–3890.

Sinha, P., Zoltners, A. A., 1979. The multiple-choice knapsack problem. Operations Research 27 (3), 503–515.

Sultan, K. S., Rahim, M. A., 1997. Optimization in quality control. Springer.

Tan, B., 1999. Variance of the output as a function of time: Production line dynamics. European Journal of Operational Research 117 (3), 470–484.

Tang, C. S., 1991. Designing an optimal production system with inspection. European Journal of Operational Research 52 (1), 45–54.

Zäpfel, G., Missbauer, H., 1993. Production planning and control (PPC) systems including load-oriented order releaseproblems and research perspectives. International Journal of Production Economics 30, 107–122.