Optimal planning of buffer sizes and inspection station positions

Mohammed Ouzineb\textsuperscript{a,b}, Fatima Zahra Mhada\textsuperscript{c}, Robert Pellerin\textsuperscript{d} and Issmail El Hallou\textsuperscript{b,d}

\textsuperscript{a}Institut National de Statistique et d’Economie Appliquee, Rabat, Morocco; \textsuperscript{b}Groupe d’etudes et de recherche en analyse des decisions, Quebec, Canada; \textsuperscript{c}ENSIAS, Universite Mohammed V de Rabat, Rabat, Morocco; \textsuperscript{d}Mathematics and Industrial Engineering, Ecole Polytechnique, Quebec, Canada

ABSTRACT
The problem of buffer sizing and inspection stations positioning in unreliable production lines is a complex mixed integer nonlinear optimization problem. In this problem, we have a production line with \( n \) machines and \( n \) fixed-size (storage) buffers in series. The machines produce parts that are either conforming or nonconforming, and the line includes inspection stations that reject the nonconforming parts. The goal is to find the optimal buffer sizes, the number and positions of the inspection stations, and satisfy the customer demand on conforming parts while minimizing the total cost. We present in this paper an exact method to solve this complex manufacturing problem. We also present new theoretical results on buffer-size bounds, stationarity, and cost function convexity permitting to significantly reduce the problem complexity. These theoretical and algorithmic developments allow solving to optimality instances with up to 30 machines tools developed previously cannot solve.

1. Introduction

A production line (automatic or not) includes storage areas called buffers. To a certain extent, buffers help in making the production stations (machines) more independent to guarantee each machine will not be affected by the unreliability of other machines (down-times, blocking, starving). To avoid such delays in manufacturing, it is important to research intermediate buffer allocation and sizing. Despite this problem, a very small number of papers address the concept of parts quality in these intermediate storage areas. Quality and quantity in manufacturing systems have long been studied separately despite their interdependence. Often, the effect of a failure due to quality on a problem formulated with respect to quantity of the production line is ignored. Both quantity and quality modeling have the same objectives: minimizing costs and maximizing productivity. Maintaining a high quality however requires early inspection for failures, which implies a reduction in inventory in the line. However, a low inventory may make the production...
line more vulnerable to machine failures and decrease its productivity. Therefore, there is a need to integrate quantity and quality modeling to optimize the performance of the production line.

Failures often occur because of a random defect in the raw material or a lack of machine maintenance, and they can be represented by independent Bernoulli random variables: a conforming part is produced with probability $f$ and a nonconforming part is produced with probability $1 - f$. This method of representing failures, i.e. following the Bernoulli distribution, is assumption in most of the literature on the inspection allocation problem (Deliman & Feldman, 1996; Han, Lim, & Park, 1998; Penn & Raviv, 2008; Raz, 1986; Viswanadham, Sharma, & Taneja, 1996).

The inspection allocation problem may include several subproblems such as the determination of (i) the optimal allocation of inspection stations within a manufacturing system (Chakravarty & Shtub, 1987; Kogan & Raz, 2002; Shiau, 2003); (ii) the optimal percentage of components to inspect (Kakade, Valenzuela, & Smith, 2004); (iii) the best action to take at an inspection station such as rework, repair, or replace (Kakade et al., 2004; Kogan & Raz, 2002; Shiau, 2003). A large part of published articles related to the inspection suppose the inspection reliable and without errors. They assumed that the produced items are subject to 100% inspection and there are no inspection errors. Some researchers have incorporated quality inspection errors caused, among others, by human failure in their models (see Hsu & Hsu, 2013a; Khan, Jaber, & Ahmad, 2014).

In a real manufacturing context, simulation can be used to determine the optimal buffers’ sizes and the optimal positions of the inspection stations by considering all possible scenarios. However, this is not practical when the systems are complex: the number of scenarios is $\binom{m}{n} \times 100^n$ for a system with $n$ machines and $m$ inspection stations where each buffer of equal size is discretized to 100 levels. Note that the system is required to have an inspection station at the end of the line. Since the position of this station is known, it is not included in the optimization. An alternative to simulation is to develop a realistic model and a fast optimization technique to solve it. The real model is however very complex because many factors impact quality and production. So, we need elaborate some simplifying assumptions in such a way that the resulting model is still realistic and can be solved optimally.

Many researchers have studied the correlation between quantity and quality, because in most real-life situations, the generation of defective parts is inevitable (see reviews by Inman, Blumenfeld, Huang, and Li (2003, 2013); Li, Blumenfeld, & Marin (2008); Mandroli, Shrivastava, & Ding (2006); Shetwan, Vitanov, & Tjahjono (2011)) and a recently overview on quality/production literature: unreliable lines (Table 1).

Kim and Gershwin (2005, 2008) studied the relationship between quality and productivity by assuming that machines can enter a failure mode that is absorbing until proper maintenance is carried out. Colledani and Tolio (2005, 2006, 2009, 2011) considered a production system composed of unreliable manufacturing and inspection stations with different failure modes. Statistical quality control charts are introduced at the inspection stations and act as noisy measurements of the quality state of the machines. Decomposition methods for studying the integrated production/quality performance of the line were developed.

In the context of serial production lines consisting of production and inspection machines that follow Bernoulli distribution failure assumptions, Meerkov and Zhang (2010)
Table 1. Overview on quality/production literature: unreliable lines.

| Article                              | Topology of the line | Keywords | Solution |
|--------------------------------------|----------------------|----------|----------|
| Assid, Gharbi, and Hajji (2015)      | PP                   | PC       | PM       | TCM      | MM       |
| Bettayeb, Bassetto, and Sahnoun (2014)| PP                   | IA       | PC       | –        | PE       | MM       |
| Bouslah, Gharbi, and Pellerin (2016) | PP 100%              | PC       | PM       | TCM      | MM       |
| Dhouib, Gharbi, and Ben Aziza (2012) | PP 100%              | PC       | PM       | TCM      | MM       |
| Gunay and Kula (2016)                | PP                   | EPQ      | –        | PrM      | MM       |
| Hsu and Hsu (2013b)                 | PP                   | IA       | PC       | –        | TCM      | CA       |
| Jafari and Makis (2015)              | PP                   | EI       | EPQ      | PM       | TCM      | CA       |
| Korytkowski and Wisniewski (2012)    | PP                   | IA       | PC       | –        | TCM      | SI       |
| Kutzner and Kiesmüller (2013)       | 1M                   | PI       | PC       | –        | TCM      | SI       |
| Lesage and Dehombreux (2015)        | PP                   | PI       | –        | PM       | TCM      | SI       |
| Liao (2013)                          | PP 2M                | –        | EPQ      | –        | PM       | TCM      | MM       |
| Matta and Simone (2016)              | PP 2M                | –        | PC       | MFM      | –        | PE       | MM       |
| Naebulharam and Zhang (2014)         | 2M 100%              | PC       | –        | PE       | MM       |
| Achille, Pellerin, and Kenne (2012)  | 1M                   | –        | PC/MFM   | PM       | TCM      | MM       |
| Nourelfath, Nahas, and Ben-Daya (2016)| 1M                   | PI       | EPQ      | PM       | TCM      | MM       |
| Rakiman and Bon (2013)               | S                    | IA       | –        | –        | PrM      | SI       |
| Raviv (2013)                         | S                    | IA       | –        | –        | PrM      | CA       |
| Rivera-Gómez, Gharbi, and Kenné (2013a)| 1M                   | –        | PC       | PM       | TCM      | SI       |
| Rivera-Gómez, Gharbi, and Kenné (2013b)| 1M                   | –        | PC/MFM   | PM       | TCM      | MM       |
| Sana (2012)                          | PP                   | –        | EPQ      | PM       | TCM      | MM       |
| Tai (2013)                           | PP                   | IE       | EPQ      | –        | TCM      | MM       |

Notes: PP: Production Process; S: Serial production line; 1M: One Machine and a stock point; 2M: Two Machines line; 100%: 100% inspection; PI: Periodic Inspection; IA: Inspection Allocation; IE: Inspection with Error; PC: Production Control; MFM: Multiple Failure Modes; EPQ: Economic Production Quality; PM: Preventive Maintenance; TCM: Total Cost Minimization; PE: Performance Evaluation; PrM: Profil Maximization; MM: Mathematical Model; CA: Computational Algorithm; SI: Simulation.

provided important insights into the nature of production and quality bottlenecks. Such systems are encountered in automotive assembly and painting operations where the downtime is relatively short and the defects are a result of uncorrelated random events.

In the same area, Mhada, Malhamé, and Pellerin (2014a) proposed an analytical model for the integrated quality and quantity control of an unreliable production line with machines producing either conforming or nonconforming parts. The line is supposed to have an inspection station at the back. So, the goal is to find the optimal number and positions of the extra inspection stations while also specifying the optimal buffer sizes, satisfying the customer demand on conforming parts and minimizing the total cost. They use a dynamic programming approach (Bertsekas, 2005) to find a balance between frequent inspection and high levels of safety stocks for a line with 10 machines and one inspection station. When there are more inspection stations, this approach becomes inefficient. Because of this limitation of the dynamic programming approach, problems solved in the literature are of limited size, not exceeding 10 machines with one inspection station.

This paper develops a more efficient exact method for a generalization of this model to the case with a variable number of inspection stations. We reformulate the problem in such a way that for a given location of inspection stations, we can use a simple and quick shortest path algorithm to find the optimal assignment of buffer sizes. We then use an exhaustive method to determine the locations for extra inspection stations and give an optimal plan for the buffer sizes and inspection stations positions. We present new theoretical results on the buffer-size bounds, stationarity, and cost function convexity. These theoretical results
help reduce the complexity of the algorithm and the number of iterations, i.e. locations to explore.

These algorithmic developments permit to solve to optimality instances with 30 machines the other tools cannot solve. Optimization techniques on approximate model, like the one proposed in this paper, would help in selecting potential scenarios/configurations for a more realistic simulation, and hence reducing the solution time without losing solution applicability in real life context. The optimization-based simulation is a powerful evaluative tool for the design and configuration of such stochastic systems.

The remainder of the paper is organized as follows. The problem formulation is given in Section 2. Some theoretical results and the proposed algorithm are presented and discussed in Section 3. The numerical results are reported in Sections 4 and 5 provides concluding remarks.

2. Problem formulation

2.1. Problem description

Figure 1 shows a homogeneous production line processing one type of parts. The line has \( n \) processing machines \( M_i \), indicated by squares, \( n \) finite-capacity storage buffers \( B_i \) for Work In Progress (WIP) inventory, indicated by circles, and \( m + 1 \) inspection stations \( SI_i \), indicated by circles inside squares (Dallery & Gershwin, 1992 and Gershwin, 1994 provide a standard description of this type of line). As will be explained below, the line contains at least one inspection station (at the end of the line) and \( m \) extra stations.

The production process produces both conforming and nonconforming items; the quality is directly affected by the reliability of the process. We therefore define a probabilistic rate \( \beta_i \) for each machine \( M_i \) that represents the proportion of nonconforming parts produced by \( M_i \).

All the machines in the line are unreliable (they can break down). Let \( \alpha_i \) be the state of machine \( M_i \): 1 if it is operational and 0 otherwise. The time between failures is random and modeled as an exponentially distributed random variable with mean \( \frac{1}{p_i} \) and the repair time is exponentially distributed with mean \( \frac{1}{r_i} \). The production line can contain inspection stations located between the buffer \( B_i \) and the provisioning point for machine \( M_{i+1} \) to reject the nonconforming parts. The use of the inspection station is captured by a binary
variable \( \lambda_i \):

\[
\begin{align*}
\lambda_i = 1 & : \text{ an inspection station is located at the exit of buffer } B_i, \\
& \text{ and all the parts processed by machine } M_{i+1} \text{ are conforming}; \\
\lambda_i = 0 & : \text{ otherwise.}
\end{align*}
\]

To guarantee that the parts delivered to the customer are conforming, we place an inspection station at the end of the line \( (\lambda_n = 1) \). The customer demand \( d \) is assumed to be constant. To ensure that the customer demand will be satisfied, we need to know the proportion of nonconforming parts \( q_i \) in each buffer \( B_i \). This depends on \( \beta_i \) and \( \lambda_j, \forall j \in [1, \ldots, i] \):

\[
q_1 = \beta_1; \quad q_i = (1 - \lambda_{i-1}) q_{i-1} + \beta_i, \quad i \geq 2.
\]

Each part must be processed on each machine \( M_i \) for a fixed time called the processing time, which is between 0 and \( k_i \) where \( k_i \) is the maximum production rate for \( M_i \). A production line in which all the machines have the same maximal production rate is called a homogeneous (or balanced) line. In a non-homogeneous (or non-balanced) line, different machines have different processing times. We remind the reader the model and cases in this paper are all homogeneous.

Buffers help to protect the line from the effects of machine failures, but they induce expensive storage costs and capital immobilization in the factory. Inspection stations play an important role in the buffer sizing since a buffer may contain nonconforming parts that represent misused production time and reduce the usefulness of the buffer and ultimately the effectiveness of the line. However, inspection stations introduce additional inspection-related costs. Therefore, the size of the buffers must be optimized to minimize the inferred costs as will be explained in the next subsection.

### 2.2. Mathematical model

Many decomposition/aggregation strategies can be used to calculate the long-term average storage cost (see Dallery and Gershwin, 1992; Han and Park, 2002; Mbihi and Malhamé, 1998; Mbihi et al., 2001; Paik et al., 2002; Sadr and Malhamé, 2004b; Sharifnia, 1988). We apply the widely used strategy of Sadr and Malhamé (2004b). It is a performance-evaluation technique that approximates a line of \( n \) machines by a set of \( n \) virtual machines \( \tilde{M}_i, i = 1, \ldots, n \). Each virtual machine is composed of machine \( M_i \) and buffer \( B_i \).

The virtual machine \( \tilde{M}_i \) is a representation of the line to machine \( M_{i+1} \), as it appears from downstream. In other words, all machines lying between the beginning of the line and \( M_{i+1} \) will be replaced by a virtual machine \( \tilde{M}_i \). This virtual machine is allowed to be in the up/down state with repair rate \( \tilde{r}_i \) and failure rate \( \tilde{p}_i \) subject to a corrected demand \( \frac{\tilde{d}_i}{\tilde{a}_i} \) (obtained by applying the averaging principle of Sadr and Malhamé, 2004b).
The values $\tilde{r}_i$, $\tilde{p}_i$, and $\tilde{d}_i$ are calculated as follows (see Sadr and Malhamé, 2004b):

$$\tilde{r}_i = \frac{\left(\tilde{r}_{i-1} - \frac{1-a_{i-1}}{a_{i-1}}\right) + p_i}{p_i \tilde{r}_{i-1} + (\tilde{r}_{i-1} - \frac{1-a_{i-1}}{a_{i-1}}) r_i} \tilde{r}_{i-1} r_i,$$

$$\tilde{p}_i = \frac{\left(\tilde{r}_{i-1} - \frac{1-a_{i-1}}{a_{i-1}}\right) + \tilde{r}_{i-1} (p_i + r_i)}{\tilde{r}_{i-1} r_i} - 1 \tilde{r}_i,$$

$$\tilde{d}_i = d \prod_{i=j}^n (1 + \lambda_j \eta_j)$$

with $\tilde{r}_1 = r_1$ and $\tilde{p}_1 = p_1$. The coefficient $a_i$ is defined to be the total WIP availability coefficient at buffer $B_i$, $1 \leq i \leq (n-1)$, i.e. $a_i$ is the probability that the inventory level of $B_i$ is positive and does not exceed the inventory capacity $z$.

This approximation allows us to use the interesting results of Hu’s model (Hu, 1995) to calculate the long-term average storage cost $T^{(i)}$ for each virtual machine $\tilde{M}_i$, $i = 1, \ldots, n$. Hu (1995) studies a production model with a single machine producing a single part type and attempting to respond to a constant demand $d$. The machine state changes in continuous time according to a homogeneous Markov process: it changes from down to up at a rate $r$ and from up to down at a rate $p$. When the machine is up, it can produce at any rate between zero and a maximum rate $k$. The storage cost per part per unit time is $c_p$. The long-term average cost $T$ associated with an arbitrary hedging point $z$ has the form

$$T(z) = \frac{\rho \left( k \left( 1 - \exp \left( -\mu (1-\rho) z \right) \right) \right) - (p + r) z \exp \left( -\mu (1-\rho) z \right)}{(p + r) (1 - \rho \exp \left( -\mu (1-\rho) z \right))}$$

with $\rho = \frac{r(k-d)}{p}$ and $\mu = \frac{p}{(k-d)}$.

The coefficient of availability of the WIP, i.e. the steady-state probability that WIP is available is

$$a = 1 - \frac{p}{(p + r) \left( 1 - \rho \exp \left( -\mu (1-\rho) z \right) \right)}.$$  

(5)

We replace the variable $z$ by $a$ to obtain another expression for the cost $T$:

$$z = \frac{1}{-\mu (1-\rho)} \ln \left[ \frac{1}{\rho} \left( 1 - \frac{(1-\rho)}{(a) \left( \frac{(p+r)}{p} \right)} \right) \right]$$

(6)

and

$$T(a) = c_p \left( \frac{k p}{\sigma (k-d) (r+p)} - \frac{k (1-a)}{\sigma (k-d)} - \frac{1}{\sigma} - \frac{(1-a) (r+p)}{\sigma^2 (k-d)} \right) \ln \left[ \frac{p d}{r (k-d)} - \frac{\sigma p d}{(p+r) r (1-a)} \right]$$

(7)
with $\sigma = \frac{(p+r) d-k r}{(k-d) d}$.

Each virtual machine $\tilde{M}_i$, $i = 1, \ldots, n$ follows the Hu’s model with $a$ set to $a_i$, $r$ set to $\tilde{r}_i$, $p$ set to $\tilde{p}_i$, $k$ set to $k_i$, and $d$ set to $\frac{d_i}{a_i}$, which allows us to express the long-term average storage cost $T^{(i)}$, $i = 1, \ldots, n$ as

$$T^{(i)} = c_p \left( \frac{k_i \tilde{p}_i}{\sigma_i (k_i - \frac{d_i}{a_i}) (\tilde{r}_i + \tilde{p}_i)} - \frac{k_i (1-a_i)}{\sigma_i (k_i - \frac{d_i}{a_i})} - \frac{1}{\sigma_i} \right) \left( \frac{(1-a_i) (\tilde{r}_i + \tilde{p}_i)}{\sigma_i^2 (k_i - \frac{d_i}{a_i})} \right) \ln \left[ \frac{\tilde{p}_i \frac{d_i}{a_i}}{\tilde{r}_i (k_i - \frac{d_i}{a_i})} - \frac{\sigma_i \tilde{p}_i \frac{d_i}{a_i}}{(\tilde{r}_i + \tilde{p}_i) \tilde{r}_i (1-a_i)} \right], \; i = 1, \ldots, n-1$$

(8)

with $\sigma_i = \frac{(\tilde{p}_i + \tilde{r}_i) \frac{d_i}{a_i} - k_i \tilde{r}_i}{(k_i - \frac{d_i}{a_i}) \frac{d_i}{a_i}}$ and

$$T^{(n)} = \rho_n \left[ c_p \left( \frac{k_n (1-\exp (-\mu_n (1-\rho_n) z_n))}{\tilde{p}_n + \tilde{r}_n (1-\rho_n) \exp (-\mu_n (1-\rho_n) z_n)} \right) - (\tilde{p}_n + \tilde{r}_n) z_n \exp (-\mu_n (1-\rho_n) z_n) \right]$$

(9)

$$\left( \frac{(\tilde{p}_n + \tilde{r}_n) (1-\rho_n) \exp (-\mu_n (1-\rho_n) z_n)}{(\tilde{p}_n + \tilde{r}_n)(1-\rho_n) \exp (-\mu_n (1-\rho_n) z_n)} \right) \ln \left[ \frac{1}{\rho_n} \left( \frac{1-\rho_n}{\mu_n (1-\rho_n)} \right) \right].$$

Note that $T^{(i)}$ is a function of $a_i, a_{i-1},$ and $\lambda$ since $\tilde{r}_i$ and $\tilde{p}_i$ are functions of $a_i$ and $a_{i-1}$, and $\tilde{d}_i$ is function of the subvector $(\lambda_1, \lambda_2, \ldots, \lambda_i)$. Here under, we refer to these quantities as $T^{(i)} = T^{(i)}(a_i, a_{i-1}, \lambda)$, for $i = 1, \ldots, n-1$ and $T^{(n)} = T^{(n)}(a_{n-1}^{\text{des}}, a_n, \lambda)$. The storage cost is then $\sum_{i=1}^n T^{(i)}(a_i, a_{i-1}, \lambda)$.

To calculate the inspection cost, we assume that all the parts are inspected. Furthermore, the inspection is assumed to be fully reliable: we assume in this paper that produced items are subject to 100% inspection with no inspection errors when there is inspection. This assumption reduces the complexity of the mathematical model and permits to study some line properties. Without this, it will be necessary to add for example the probability of classifying a non-defective item as defective, the probability of classifying a defective item as non-defective, and the sample size to be inspected which complicate the theoretical study a lot that is already very difficult. The average cost per time unit caused by the existence of nonconforming parts is zero if the inspection station is located before $B_i$. Otherwise, it is proportional to the long-term average number of parts $d_i$ pulled per unit of time from buffer $B_i$ and to the inspection cost $c_l$ per pulled part. Therefore, the total inspection cost is $c_l \sum_{i=1}^n \lambda_i \tilde{d}_i$.

Finally, the total cost to be minimized is the sum of the average long-term combined storage cost and the inspection cost (see Mhada et al., 2014a):

$$J(a, \lambda) = \sum_{i=1}^n T^{(i)}(a_i, a_{i-1}, \lambda) + c_l \sum_{i=1}^n \lambda_i \tilde{d}_i.$$

(10)
We refer to the vector \((a_1, a_2, \ldots, a_n)\) by \(a\). The optimization problem is then to minimize the long-term average storage and inspection costs per unit time when \(m\) extra inspection stations are added to the line \(\sum_{i=1}^{n-1} \lambda_i = m\) and to find the best locations for these extra stations. We must ensure that 1) the WIP availability coefficient \(a_n\) at buffer \(B_n\), i.e. the probability that the WIP is active (the inventory is positive or zero when the previous machine is functional), is equal to a given availability coefficient of (conforming) finished parts \(a_{n \text{des}}\); and 2) the inventory of finished parts satisfies the demand \(d\). Mathematically speaking, the objective is to find the minimal average global system cost structure \((a, \lambda)\) such that:

\[
\begin{align*}
\text{minimize} & \quad J(a, \lambda) \\
\text{subject to} & \quad \sum_{i=1}^{n-1} \lambda_i = m \\
& \quad \lambda_i \in \{0, 1\} \; \forall i, 1 \leq i \leq n - 1 \\
& \quad a_i \geq 0 \; \forall i, 1 \leq i \leq n
\end{align*}
\]

Note that the theoretical results in Section 3 and the numerical results of Section 4 apply to homogeneous and partially homogeneous production lines: \(r_1 = r_2 = \ldots = r\) and \(\beta_1 = \beta_2 = \ldots = \beta\).

### 3. Algorithm and theoretical insights

The problem defined by (11)–(14) is an mixed integer nonlinear optimization problem (MINLP). Solving small instances of this problem by standard dynamic programming takes hours, as will be shown in Section 4. Dynamic programming cannot solve actually larger instances due to the limitation of the memory. We therefore propose an alternative algorithm. Observe that (11)–(14) is equivalent to \(\min_\lambda (\min_a J(a, \lambda))\) where \(a\) and \(\lambda\) have to satisfy constraints (12)–(14). As claimed by Proposition 1, we reformulate \(\min_a J(a, \lambda)\) for each fixed location \(\lambda\) as a shortest path problem on a network and solve it via a standard shortest path algorithm (Dijkstra’s algorithm).

Consider the connected network \(G(N, A)\) consisting of a set of nodes \(N\) and a set of arcs \(A\), as depicted in Figure 2. With each machine we associate a set of nodes representing all the possible buffers’ sizes, which range from 1 to 100%. The set \(N\) consists of \(100 \times (n-1)\) nodes (representing the \(n-1\) machines) plus two additional nodes \(N_0\) and \(N_n\) that are adjacent to the nodes representing the first and last machines. The set \(A\) consists of \(100^2 \times (n-2) + 200\) arcs weighted as follows:

- \(N_0\) is connected to the nodes \(N^j_1, j = 1, \ldots, 100\) representing machine 1. The weight of each arc \((N_0, N^j_1)\) is null. \(N_n\) is connected to the nodes \(N^{j}_{n-1}, j = 1, \ldots, 100\) representing machine \(n-1\). The weight of each arc \((N^j_{n-1}, N_n)\) is \(c_{n-1, n} = T_F(a_{n \text{des}}^j, j, \lambda) + c_I \times \tilde{d}_n, 1 \leq j \leq 100\), where \(\tilde{d}_n\) is the long-term average number of parts pulled per time unit from stock \(x_n\), \(c_I \times \tilde{d}_n\) is the inspection cost and \(T_F(a_{n \text{des}}^j, j, \lambda)\) is the long-term average...
storage cost that depends on the total work-in-progress availability coefficient $a_j$ at buffer $B_{n-1}$.

- Each node $N^j_i$ representing machine $i$ is connected to each node $N^k_{i+1}$ representing machine $i+1$ by an arc $(N^j_i, N^k_{i+1})$ that is weighted $c^{jk}_{i,i+1} = T^{(i)}(k,j,\lambda) + c_I \times \lambda_i \times \tilde{d}_i$, $i = 1, \ldots, n-2$ and $1 \leq j, k \leq 100$, where $\tilde{d}_i$ is the long-term average number of parts pulled per unit of time from stock $B_i$, $c_I \times \lambda_i \times \tilde{d}_i$ is the inspection cost and $T^{(i)}(k,j,\lambda)$ is the long-term average storage cost that depends on the total work-in-progress availability coefficient $a_j$ at buffer $B_{i-1}$.

**Proposition 1:** For each fixed location $\lambda$, the minimum total storage and inspection cost is the sum of the arc costs on the shortest path in $G$ between $N_0$ and $N_n$.

**Proof:** As mentioned before, the problem (11)–(14) is equivalent to $\min_{\lambda}(\min_a J(a, \lambda))$. Let $\mathcal{P}$ be the set of paths in $G(N,A)$ from $N_0$ to $N_n$ and $P$ a path in $\mathcal{P}$. Let $Y$ be decision variables such that

\[
Y^{jk}_{i,i+1} = \begin{cases} 1 & \text{if arc } (N^j_i, N^k_{i+1}) \in P \\ 0 & \text{otherwise;} \end{cases}
\]

\[
Y^k_{0,1} = \begin{cases} 1 & \text{if arc } (N_0, N^k_1) \in P \\ 0 & \text{otherwise;} \end{cases}
\]

\[
Y^j_{n-1,n} = \begin{cases} 1 & \text{if arc } (N^j_{n-1}, N_n) \in P \\ 0 & \text{otherwise.} \end{cases}
\]

For each location $\lambda$, we have:

\[
\min J(a, \lambda) = \min \left( \sum_{i=1}^{n-1} T^{(i)}(a_{i+1}, a_i, \lambda_0) + c_I \sum_{i=1}^{n-1} \lambda_i \tilde{d}_i + T_F(a_{n-1}^{\text{des}}, a_{n-1}, \lambda_0) + c_I \lambda_n \tilde{d}_n \right)
\]

\[
= \min \left( \sum_{i=1}^{n-2} \sum_{j=1}^{100} \sum_{k=1}^{100} c^{jk}_{i,i+1} Y^{jk}_{i,i+1} + \sum_{j=1}^{100} c^{j}_{n-1,n} Y^{j}_{n-1,n} \right)
\]  

(15)
subject to (16)

\[ \sum_{k=1}^{100} Y_{0,1}^k = 1 \]  \hspace{1cm} (17)

\[ \sum_{j=1}^{100} Y_{n-1,n}^j = 1 \]  \hspace{1cm} (18)

\[ \sum_{j=1}^{100} Y_{i-1,i}^j = \sum_{k=1}^{100} Y_{i,i+1}^k \forall l \in \{1, \ldots, 100\}, i \in \{2, 3, \ldots, n - 2\}. \]  \hspace{1cm} (19)

Observe that Equations (17)–(19) are conservation flow constraints. Consequently, the problem can be considered as a single-source shortest path problem. To find the minimum, we must find the shortest (lowest cost) path from \( N_0 \) to \( N_n \).

Proposition 1 leads to the development of Algorithm 1. Observe that there is no randomness in the behavior of the proposed exact algorithm. We have the following remark.

**Remark 1:** complexity of the new algorithm depends only on the size of the network \( G \) (defined by the number of machines and buffer levels (how we discretize the buffers)) and the number of inspections stations.

There are two ways to improve this algorithm. First, we could restrict the possible values of the buffer availability coefficient \( (a_i) \), leading to a reduction in the network size and therefore a reduction of the time per iteration. Second, we could restrict the possible positions of the inspection stations, reducing the number of iterations.

---

**Algorithm 1 Optimal buffer sizing and inspection station location**

1. \( Z^* = \infty, k = 0 \).
2. Construct the network \( G \) (with arc costs set to 0).
3. for all \( \lambda \) do
   4. Compute the arc costs for the current location \( \lambda \).
   5. Find a shortest path in \( G \) (the buffer size vector \( a \)) and set \( Z \) to its cost.
   6. if \( Z < Z^* \) then
      7. \( Z^* = Z, a^* = a, \lambda^* = \lambda \).
   8. end if
   9. \( k = k + 1 \).
4. end for
5. return \( (Z^*, a^*, \lambda^*) \).

---

To reduce the number of nodes and arcs in \( G(N,A) \), we use characteristics of the production line to prove theoretical results on the possible values of \( a_i \).

**Lemma 1:** The coefficient of availability \( a_i \) of the buffer \( B_i \) has a lower bound of \( \frac{r}{r+p_i} a_{i-1} \).

**Proof:** Buffer \( B_i \) will contain parts if machine \( M_i \) is functional and \( B_{i-1} \) contains parts, i.e. \( P[x_i \geq 0] \geq P[x_{i-1} \geq 0, M_i \text{ is ON}] \).

By definition \( a_i = P[x_i \geq 0], P[x_{i-1} \geq 0, M_i \text{ is ON}] = P[x_{i-1} \geq 0]P[M_i \text{ is ON}], \) and \( P[M_i \text{ is ON}] = \frac{r}{r+p_i} \). Consequently, \( a_i \geq \frac{r}{r+p_i} a_{i-1} \). \( \square \)
Proposition 2:  The buffer size $a_i$ has an upper bound of $\min\left(\prod_{j=i}^{n} \left[ \frac{r + p_j}{r} \right] a_n^{\text{des}}, 1 \right)$.

Proof: From Lemma 1, we have

$$a_i \geq \frac{r}{r + p_i} a_{i-1} \Rightarrow a_{i-1} \leq \frac{r + p_i}{r} a_i.$$ 

The recurrence of this equation leads us to conclude that $a_{i-1} \leq \prod_{j=i}^{n} \left[ \frac{r + p_j}{r} \right] a_n^{\text{des}}$, and by definition $a_n = a_n^{\text{des}}$, so $a_{i-1} \leq \prod_{j=i}^{n} \left[ \frac{r + p_j}{r} \right] a_n^{\text{des}}$. $\square$

Lemma 2:  The coefficient of availability $a_i$ of buffer $B_i$ has a lower bound of $\frac{r + p_i}{r k_i} d$.

Proof: The second characteristic of the production line is that each machine $M_i$ is able to meet the demand $d$. Thus, the production rate of $M_i$, when it is functional and has enough parts to process, must be at least $d$, i.e.

$$k_i P[M_i \text{ is ON}] \geq d \Rightarrow k_i \frac{r}{r + p_i} a_{i-1} \geq d \Rightarrow a_{i-1} \geq \frac{r + p_i}{r k_i} d.$$ 

$\square$

Proposition 3:  The buffer size $a_i$ has a lower bound of $\max\left(\prod_{j=1}^{i-1} \left[ \frac{r}{r + p_j} \right], \frac{r + p_i}{r k_i} d \right)$.

Proof: From Lemma 1, we have

$$a_i \geq \frac{r}{r + p_i} a_{i-1}.$$ 

The recurrence of this equation leads us to conclude that:

$$a_{i-1} \geq \prod_{j=1}^{i-1} \left[ \frac{r}{r + p_j} \right] a_0.$$ 

Here $a_0$ is the raw material availability coefficient, and since we assume that raw material will always be available, $a_0 = 1$ and

$$a_{i-1} \geq \prod_{j=1}^{i-1} \left[ \frac{r}{r + p_j} \right].$$ 

(20)

From Lemma 2, we have

$$a_{i-1} \geq \frac{r + p_i}{r k_i} d.$$
Hence,

\[ a_{i-1} \geq \max \left( \prod_{j=1}^{i-1} \left[ \frac{r}{r + p_j}, \frac{r + p_i}{r k_i} d \right] \right). \]

The bounds of Propositions 2 and 3 reduce the number of nodes and arcs in \( G \) (Figure 2). Each column \( i \) contains only nodes with values between \( 100 \times \max \left( \prod_{j=1}^{i-1} \left( \frac{r}{r + p_j} \right) \right) \) and \( 100 \times \min \left( \prod_{j=i}^{n} \left( \frac{r + p_j}{r} a_{\text{des}}^{n} \right) \right) \). Nodes with values outside this range are removed. Clearly, a quadratic number of arcs between the removed nodes are also eliminated. To further reduce the solution space, we apply a stationary characteristic of homogeneous lines. We give an example in Section 4.4 to illustrate the theoretical result below. This result is based on the interesting Proposition 2 of Sadr and Malhamé (2004a), which is Proposition 4 below.

Let

\[ P_s : \quad \min_a J(a, \lambda_0) = \min_a \sum_{i=1}^{n-1} T^{(i)}(a_{i-1}, a_i, \lambda_0) \]

with

\[ T^{(i)}(a_{i-1}, a_i, \lambda_0) = c_p \left( \frac{k (1 - a_i) \tilde{d}_i}{a_i} \right) - \frac{k (1 - a_i) \tilde{d}_i}{a_i} - \frac{\left( p+r(1-a_{i-1}) \right) \tilde{d}_i}{a_i} - \frac{\left( p+r(1-a_{i-1}) \right) \tilde{d}_i}{a_i} - \frac{\left( p+r(1-a_{i-1}) \right) \tilde{d}_i}{a_i} - \frac{\left( p+r(1-a_{i-1}) \right) \tilde{d}_i}{a_i} \]

\[ \ln \left( \frac{p+r(1-a_{i-1}) \tilde{d}_i}{a_i} - \frac{\left( p+r(1-a_{i-1}) \right) \tilde{d}_i}{a_i} - \frac{\left( p+r(1-a_{i-1}) \right) \tilde{d}_i}{a_i} \right), \]

\[ i = 1, \ldots, n - 1. \]

Sadr and Malhamé (2004a) show the following result.

**Proposition 4:** For a perfectly homogeneous line (identical \( r_i, p_i, k_i, \text{and} \beta_i \)), as the number of machines \( n \) goes to infinity, the problem \( P_s \) admits a stationary-state feedback control policy that is optimal. This policy contains some constant availability coefficients \( \tilde{a} = \arg \min T^{(i)}(a, a, 0), \) subject to \( \max \left[ \left( \frac{r + p}{r + p} \right)^i, \frac{\tilde{d}_i (r + p)}{k} \right] < a < 1. \)

This allows us to establish the result below. The idea is to extend Proposition 4, relating to a homogeneous line without any inspection station (\( m = 0 \)), to lines with \( m > 0 \). This new result is confirmed numerically in Section 4. It could be particularly useful to find an initial solution that is close to the optimum for very large lines.
**Proposition 5:** For a perfectly homogeneous line (with identical $r_i$, $p_i$, $k_i$, and $\beta_i$) with a sufficiently large number of machines ($n$ goes theoretically to infinity) and a fixed $\lambda = \lambda_0$, there is a stationary feedback control policy that is optimal. This policy contains some constant availability coefficients $\bar{a}$ for the machines considered ‘sufficiently far’ from the first machine, meaning that the optimal profile will be flat at the fixed level $\bar{a}$, independent of the boundary condition on $a_0$ and $a_n$.

**Proof:** The idea of the proof is to extend Proposition 4 to lines with $m > 0$ inspection stations in the positions $e_1$, $e_2$, ..., $e_m$ with $e_m < n$. We therefore divide the line into $m + 1$ homogeneous lines separated by $m$ inspection stations. The parameters of these lines, except for the pulled demand $\tilde{d}_i$, are identical since

$$
\begin{align*}
1 \leq i \leq e_1 & \quad q_i = (1 + \beta)^i - 1 \quad \tilde{d}_i = d (1 + \beta)^n \\
e_1 < i \leq e_2 & \quad q_i = (1 + \beta)^{i-e_1} - 1 \quad \tilde{d}_i = d (1 + \beta)^{n-e_1} \\
\vdots & \quad \vdots \\
e_{m-1} < i \leq e_m & \quad q_i = (1 + \beta)^{i-e_{m-1}} - 1 \quad \tilde{d}_i = d (1 + \beta)^{n-e_{m-1}}.
\end{align*}
$$

We then apply Proposition 4 to each line segment with $\bar{a} = \arg\min_{T(i)} T(a, a, \lambda_0)$. This allows us to calculate the values of the optimal availability coefficients in which each line segment will be flat.

Conjectures 1 and 2 emphasize the fact that the cost function is convex in the position of the inspection station and the number of inspection stations. These conjectures are likely to be true according to our numerical results. The proofs of these conjectures are left as open questions.

**Conjecture 1:** The minimal total cost (storage and inspection costs) is a convex function of $\lambda$, i.e. the location of the internal inspection station.

**Conjecture 2:** The minimal total cost (storage and inspection costs) is a convex function of the number of internal inspection stations.

These conjectures about the convexity of the cost function help reducing the number of iterations of the algorithm and therefore the solution time. For instance, the ‘local minimum’ of a convex function is also a global minimum, and there are many efficient methods for optimizing convex functions.

4. **Numerical results**

To test the efficiency of the exact method (EM) (presented by Algorithm 1), we compared it to dynamic programming (DP) approach on five line production lines ranging from a small one with 10 machines to a large one with 30 machines. DP is usually used to solve such problems (Mhada, Malhamé, & Pellerin, 2014b). Note please that these problems are unsolvable with generic nonlinear programming solvers like MINOS. We implemented EM in C++. The tests were performed on a computer (Desktop) with a Quad Core Intel i7 processor clocked at 2.8 GHz per core, 8 GB of system memory, and running a Linux operating system.
Table 2. System parameters.

|                  | 10 machines | 15 machines | 20 machines | 25 machines | 30 machines |
|------------------|-------------|-------------|-------------|-------------|-------------|
| $n$              | 10          | 15          | 20          | 25          | 30          |
| $p_i$            | 2           | 2           | 2           | 2           | 2           |
| $r_i$            | .9          | .9          | .9          | .9          | .9          |
| $k_i$            | 4           | 7           | 9           | 14          | 22          |
| $\beta_i$       | .1          | .1          | .1          | .1          | .1          |
| $d$              | 1           | 1           | 1           | 1           | 1           |
| $c_p$            | 1           | 1           | 1           | 1           | 1           |
| $c_I$            | 2           | 2           | 2           | 2           | 2           |
| $a_{des}^n$      | .95         | .95         | .95         | .95         | .95         |

Figure 3. Optimal cost as a function of $\lambda_i$, $i = 1, 2, \ldots, 9$: 10 machines and $m = 1$.

4.1. Benchmark data

Table 2 lists the parameters used. For machine $M_i (i = 1, 2, \ldots, n)$, $k_i$ is the maximal production rate, $p_i$ the failure rate, $r_i$ is the repair rate, $\beta_i$ is the ratio of nonconforming parts to conforming parts, $d$ is the demand, $c_p$ is the storage cost per time unit and per part, $c_I$ is the inspection cost per pulled part, $m$ is the number of inspection stations, and $a_{des}^n$ is the required availability coefficient of finished parts. The test instances are available upon request.

4.2. Optimal location of inspection station for $m = 1$

Figures 3 and 4 display the optimal cost as a function of the location of the internal inspection station $\lambda_i$, $i = 1, 2, \ldots, (n - 1)$ for two test problems (10 machines and 20 machines). We observe that in our homogeneous production line the minimal total cost is a convex function of the location of the internal inspection station $i$, i.e. $\lambda_i = 1$. The optimal position is at position four (i.e. $\lambda_4 = 1$) for both problems. This convexity is also observed on the other lines. This observation confirms Conjecture 1. This is an intuitive result because if the station is placed at the beginning of the line it will be less effective at rejecting nonconformities, and if it is placed at the end there will be more nonconformity during the fabrication process and thus a high storage cost.
4.3. Cost function according to the number of inspection stations

Figures 5–9 display the optimal cost as a function of \(m\) (i.e. the number of inspection stations for the case \(m \geq 2\)) for the five test problems. The minimal total cost is a convex function of the number of inspection stations (on all lines). This confirms Conjecture 2. For instance, the optimal number is one (inspection) station for the 10 machines line and three for the 20 machines. In the latter case, the optimal stations positions are at positions 2, 7 and 18 (\(\lambda_2 = \lambda_7 = \lambda_{18} = 1\)). Figure 10 displays the optimal availability coefficients \(a_i\) for each buffer \(B_i, i = 1, \ldots, (n - 1)\) for the third test.
4.4. Availability coefficient state space

For the 20 machines line, with four inspection stations placed after \( M_2, M_7, M_{18}, \) and \( M_{20} \), we calculate \( \bar{a} \) for each line segment located between two consecutive stations that can be considered to contain a sufficiently large number of machines. We obtain

- For the line segment from \( M_3 \) to \( M_7 \): \( \bar{a} = .7649 \) (Bold values in Table 3 for \( i=2..6 \)).
- For the line segment from \( M_8 \) to \( M_{18} \): \( \bar{a} = .4758 \) (Bold values in Table 3 for \( i=9..17 \)).

We compare these \( \bar{a} \)'s with the availability coefficients of the optimal solution found by the exhaustive method (Table 3 and Figure 10), and we note that it is possible to improve the running time. The availability coefficients \( a_i^0, i = 1, \ldots, n - 1 \) for this initial solution
The optimal cost can be computed as follows:

\[ a_i^0 = \max \left[ \bar{a}, \max \left( \left( \frac{r}{r + p} \right)^{(i)}, \frac{\tilde{d}_i}{k} \frac{r + p}{r} \right) + .01 \right] \]

These numerical results are coherent with Proposition 5.
The position of the machine $M_i$

The optimal availability coefficient $a_i$

Figure 10. Optimal availability coefficient: 20 machines.

Table 3. Optimal availability coefficient state space $a_i, i = 1, \ldots, n - 1$: 20 machines.

| $i$ | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|-----|----|----|----|----|----|----|----|----|----|----|
| $a_i$ | .92 | .77 | .77 | .77 | .77 | .76 | .63 | .52 | .48 | .48 |
| $i$ | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| $a_i$ | .48 | .48 | .48 | .48 | .48 | .47 | .39 | .54 | .95 |    |

4.5. Comparison with dynamic programming

To fairly compare EM to DP, we implemented the latter under the same conditions as for EM (computer, programming language, operating system). For the small instances with 10 machines, DP is competitive against EM. Tables 4 and 5 give the results of the comparison for midsize instances (15 and 20 machines). These results show that EM generally outperforms DP. DP is slightly better for very small $m$ ($m \leq 2$). EM is largely better for bigger $m$. EM consistently uses less memory than DP. Used memory in DP increases significantly with $m$ while it is constant in EM which is very interesting. Actually, DP runs out of memory on 20 machines instances with $m \geq 8$.

Table 6 shows that DP is unable to solve any instance with 25 machines due to memory issues. Opposite to that, EM consumes a stable and small amount of memory and is able to solve to optimality these large instances.

For the 30 machines instances (see Table 7), EM finds an optimal solution in a time that could be considered reasonable (around 4 days if solved in parallel) for a design problem. The solution is optimal for $m = 5$ (minimum cost) thanks to convexity. We report in this table also the optimal costs for these instances. DP runs out of memory for all values of $m$. However, EM takes too much time to solve the instances with $m \geq 7$. For example, for $m = 7$, EM converges after a month of running time. Large problems require fast heuristics that could provide good results in an acceptable time.
Table 4. Comparison of EM and DP for 15 machines.

| m   | EM (s) | DP (s) | EM (Mb) | DP (Mb) |
|-----|--------|--------|---------|---------|
| 1   | 2.57   | 4.54   | 30      | 514     |
| 2   | 16.81  | 38.18  | 30      | 565     |
| 3   | 67.28  | 175.95 | 30      | 719     |
| 4   | 190.77 | 561.08 | 30      | 873     |
| 5   | 384.56 | 1183.67| 30      | 1079    |
| 6   | 577.2  | 1819.31| 30      | 1233    |
| 7   | 793.72 | 2102.85| 30      | 1387    |
| 8   | 698.33 | 1895.68| 30      | 1542    |
| 9   | 481.08 | 1287.155| 30   | 1696    |
| 10  | 212.36 | 655.05 | 30      | 1799    |

Table 5. Comparison of EM and DP for 20 machines.

| m   | EM (s) | DP (s) | EM (Mb) | DP (Mb) |
|-----|--------|--------|---------|---------|
| 1   | 15     | 7      | 52      | 564     |
| 2   | 105    | 89     | 52      | 628     |
| 3   | 421    | 563    | 52      | 894     |
| 4   | 1601   | 2610   | 52      | 1022    |
| 5   | 4743   | 7651   | 52      | 1291    |
| 6   | 11108  | 18394  | 52      | 1534    |
| 7   | 20613  | 33846  | 52      | 2195    |
| 8   | 29790  | –      | 52      | out of memory |
| 9   | 35764  | –      | 52      | out of memory |
| 10  | 35767  | –      | 52      | out of memory |

Table 6. Comparison of EM and DP for 25 machines.

| m   | EM (s) | DP (s) | EM (Mb) | DP (Mb) |
|-----|--------|--------|---------|---------|
| 1   | 12.9   | –      | 90      | out of memory |
| 2   | 150.4  | –      | 90      | out of memory |
| 3   | 258.4  | –      | 90      | out of memory |
| 4   | 6252.5 | –      | 90      | out of memory |
| 5   | 25794.0| –      | 90      | out of memory |
| 6   | 74452.4| –      | 90      | out of memory |
| 7   | 183610 | –      | 90      | out of memory |

Table 7. Results for 30 machines.

| m   | Cost      | Running time (s) |
|-----|-----------|------------------|
| 1   | 40.7297   | 31.2             |
| 2   | 22.9847   | 389.4            |
| 3   | 20.0887   | 3524.2           |
| 4   | 19.0494   | 18769.2          |
| 5   | 18.9133   | 115432.0         |
| 6   | 19.1256   | 421630.0         |
5. Conclusions

We present in this paper an exact method to solve a complex manufacturing problem formulated as a MINLP with more than 2000 variables in some cases. Dynamic programming and generic nonlinear programming solvers cannot solve reasonable size instances of this problem. We develop some theoretical results that help in significantly reducing the problem complexity. We show that the objective function is convex in the position of the inspection station and in the number of internal inspection stations. We also reformulate the problem, for a given location of inspection stations, as a shortest path problem to rapidly find the optimal buffer sizes. These new theoretical and algorithmic developments permit to solve to optimality instances with up to 30 machines the tools developed previously cannot solve.

Future research should focus on developing an optimization method for more realistic non-homogeneous production lines. It would be interesting to demonstrate if the convexity (Figures 5–9) remains valid in the non-homogeneous case. We also plan to combine our method with a metaheuristic to reduce computational time without sacrificing the solutions’ quality provided by the exact method or to combine it with simulation. The model developed here could be used as an approximation model for more complex real life design problems. Fast optimization techniques, like the one proposed in this paper or EM based heuristic approach, would help in rapidly selecting potential scenarios/configurations for a more realistic simulation, and hence reducing the solution time without losing solution applicability in real life context.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Achille, N., Pellerin, R., & Kenne, J. (2012). Simultaneous control of maintenance and production rates of a manufacturing system with defective products. *Journal of Intelligent Manufacturing, 23*(2), 323–332.

Assid, M., Gharbi, A., & Hajji, A. (2015). Joint production, setup and preventive maintenance policies of unreliable two-product manufacturing systems. *International Journal of Production Research, 53*(15), 4668–4683.

Bertsekas, D. (2005). *Dynamic programming and optimal control* (3rd ed.). Belmont, MA: Athena Scientific.

Bettayeb, B., Bassetto, S., & Sahnoun, M. (2014). Quality control planning to prevent excessive scrap production. *Journal of Manufacturing Systems, 33*(3), 400–411.

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.

Chakravarty, A. K., & Shtub, A. (1987). Strategic allocation of inspection effort in a serial, multi-product production system. *IIE Transactions, 19*(1), 13–22.

Colledani, M., & Tolio, T. (2005). *Impact of statistical process control (spc) on the performance of production systems-part 2 (large systems)*. Greece: Zakynthos Island.

Colledani, M., & Tolio, T. (2006). Impact of quality control on production system performance. *Annals of the CIRP, 55*(1), 453–456.

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

Dallery, Y., & Gershwin, S. (1992). Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems, 12*, 3–94.

Deliman, N., & Feldman, R. (1996). Optimization of process improvement and inspection location for serial manufacturing. *International Journal of Production Research, 34*(2), 395–405.

Dhouib, K., Gharbi, A., & Ben Aziza, M. (2012). Joint optimal production control/preventive maintenance policy for imperfect process manufacturing cell. *International Journal of Production Economics, 137*(1), 126–136.

Gershwin, S. (1994). *Manufacturing systems engineering*. Upper Saddle River, NJ: Prentice Hall.

Gunay, E., & Kula, U. (2016). Integration of production quantity and control chart design in automotive manufacturing. *Computers & Industrial Engineering, 102*, 374–382.

Han, M., Lim, J., & Park, D. (1998). Performance analysis of serial production lines with quality inspection machines. *International Journal of System Science, 29*(9), 939–951.

Han, M.-S., & Park, D.-J. (2002). Optimal buffer allocation of serial production lines with quality inspection machines. *Computers & Industrial Engineering, 42*(1), 75–89.

Hsu, J.-T., & Hsu, L.-F. (2013a). An EOQ model with imperfect quality items, inspection errors, shortage backordering, and sales returns. *International Journal of Production Economics, 143*(1), 162–170.

Hsu, J.-T., & Hsu, L.-F. (2013b). Two EPQ models with imperfect production processes, inspection errors, planned backorders, and sales returns. *Computers & Industrial Engineering, 64*(1), 389–402.

Hu, J. (1995). Production rate control for failure-prone production systems with no backlog permitted. *IEEE Transactions on Automatic Control, 40*(2), 291–295.

Inman, R., Blumenfeld, D., Huang, N., & Li, J. (2003). Designing production systems for quality: Research opportunities from automotive industry prospective. *International Journal of Production Research, 41*(9), 1953–1971.

Jafari, L., & Makis, V. (2015). Joint optimal lot sizing and preventive maintenance policy for a production facility subject to condition monitoring. *International Journal of Production Economics, 169*, 156–168.

Kakade, V., Valenzuela, J., & Smith, J. (2004). An optimization model for selective inspection in serial manufacturing systems. *International Journal of Production Research, 42*(18), 3891–3909.

Khan, M., Jaber, M. Y., & Ahmad, A.-R. (2014). An integrated supply chain model with errors in quality inspection and learning in production. *Omega, 42*(1), 16–24.

Kim, J., & Gershwin, S. (2005). Integrated quality and quantity modeling of a production line. *Stochastic Modeling of Manufacturing Systems, 27*(2–3), 287–314.

Kim, J., & Gershwin, S. (2008). Analysis of long flow lines with quality and operational failures. *IIE Transactions, 40*(3), 284–296.

Kogan, K., & Raz, T. (2002). Optimal allocation of inspection effort over a finite planning horizon. *IIE Transactions, 34*(6), 515–527.

Korytkowski, P., & Wisniewski, T. (2012). Simulation-based optimisation of inspection stations allocation in multi-product manufacturing systems. *International Journal of Advanced Operations Management, 4*(1/2), 105–123.

Kutzner, S. C., & Kiesmüller, G. P. (2013). Optimal control of an inventory-production system with state-dependent random yield. *European Journal of Operational Research, 227*(3), 444–452.

Lesage, A., & Dehombreux, P. (2015). Simulation of manufacturing processes subject to maintenance and quality control: Sensitivity analysis and numerical example. *International Journal of Mechanical Engineering and Automation, 2*(2), 119–126.

Li, J., Blumenfeld, D., & Marin, S. P. (2008). Production system design for quality robustness. *IIE Transactions, 40*(3), 162–176.

Liao, G.-L. (2013). Optimal economic production quantity policy for randomly failing process with minimal repair, backorder and preventive maintenance. *International Journal of Systems Science, 44*(9), 1602–1612.
Mandroli, S. S., Shrivastava, A. K., & Ding, Y. (2006). A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes. *IIE Transactions*, 38(4), 309–328.

Matta, A., & Simone, F. (2016). Analysis of two-machine lines with finite buffer, operation-dependent and time-dependent failure modes. *International Journal of Production Research*, 54(6), 1850–1862.

Mbihi, J., & Malhamé, R. (1998). *Optimization of a class of decentralized hedging production policies in an unreliable two-machine flow shop*. FL: Tampa.

Mbihi, J., Malhamé, R., & Sadr, J. (2001). *Two approximations as a basis for the optimization of production in long unreliable markovian transfer lines* (ed.). Amsterdam: Kluwer.

Meerkov, S., & Zhang, L. (2010). Product quality inspection in bernoulli lines: analysis, bottlenecks, and design. *International Journal of Production Research*, 48(16), 4745–4766.

Mhada, F., Malhamé, R., & Pellerin, R. (2014a). Joint assignment of buffer sizes and inspection points in unreliable transfer lines with scrapping of defective parts. *Production and Manufacturing Research*, 1(1), 79–101.

Mhada, F., Malhamé, R., & Pellerin, R. (2014b). A stochastic hybrid state model for optimizing hedging policies in manufacturing systems with randomly occurring defects. *Discrete Event Dynamic Systems*, 24, 69–98.

Naebulharam, N., & Zhang, L. (2014). Bernoulli serial lines with deteriorating product quality: performance evaluation and system-theoretic properties. *International Journal of Production Research*, 52(5), 1479–1494.

Nourelfath, M., Nahas, N., & Ben-Daya, M. (2016). Integrated preventive maintenance and production decisions for imperfect processes. *Reliability Engineering & System Safety*, 148, 21–31.

Paik, C., Kim, H., & Cho, H. (2002). Performance analysis for closed-loop production systems with unreliable machines and random processing times. *Computers & Industrial Engineering*, 42(2–4), 207–220.

Penn, M., & Raviv, T. (2008). A polynomial time algorithm for solving a quality control station configuration problem. *Discrete Applied Mathematics*, 156(4), 412–419.

Rakiman, U. S. B., & Bon, A. T. (2013). Production line: Effect of different inspection station allocation. *Procedia Engineering*, 53, 509–515.

Raviv, T. (2013). An efficient algorithm for maximizing the expected profit from a serial production line with inspection stations and rework. *OR Spectrum*, 35(3), 609–638.

Raz, T. (1986). A survey of models for allocating inspection effort in multistage production systems. *Journal of quality technology*, 18(4), 239–246.

Rivera-Gómez, H., Gharbi, A., & Kenné, J. P. (2013a). Joint production and major maintenance planning policy of a manufacturing system with deteriorating quality. *International Journal of Production Economics*, 146(2), 575–587.

Rivera-Gómez, H., Gharbi, A., & Kenné, J. P. (2013b). Production and quality control policies for deteriorating manufacturing system. *International Journal of Production Research*, 51(11), 3443–3462.

Sadr, J., & Malhamé, R. (2004a). Decomposition/aggregation-based dynamic programming optimization of partially homogeneous unreliable transfer lines. *IEEE Transactions on Automatic Control*, 49(1), 68–81.

Sadr, J., & Malhamé, R. (2004b). Unreliable transfer lines: decomposition/ aggregation and optimisation. *Annals of Operations Research*, 125(1–4), 167–190.

Sana, S. S. (2012). Preventive maintenance and optimal buffer inventory for products sold with warranty in an imperfect production system. *International Journal of Production Research*, 50(23), 6763–6774.

Sharifnia, A. (1988). Production control of a manufacturing system with multiple machine states. *IEEE Transactions on Automatic Control*, 33, 620–625.

Shetwan, A. G., Vitanov, V. I., & Tjahjono, B. (2011). Allocation of quality control stations in multistage manufacturing systems. *Computers & Industrial Engineering*, 60(4), 473–484.

Shiau, Y. (2003). Inspection allocation planning for a multiple quality characteristic advanced manufacturing system. The *International Journal of Advanced Manufacturing Technology*, 21(7), 494–500.
Tai, A. H. (2013). Economic production quantity models for deteriorating/imperfect products and service with rework. *Computers & Industrial Engineering, 66*(4), 879–888.

Viswanadham, N., Sharma, S., & Taneja, M. (1996). Inspection allocation in manufacturing systems using stochastic search techniques. *IEEand Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 26*(2), 222–230.