Optimal Sampling Plan for an Unreliable Multistage Production System Subject to Competing and Propagating Random Shifts

SINAN OBAIDAT and HAITAO LIAO*

Department of Industrial Engineering, University of Arkansas, Fayetteville, AR 72701, USA

E-mail: {sfobaida, liao}@uark.edu

Abstract

Sampling plans play an important role in monitoring production systems and reducing quality- and maintenance-related costs. Existing sampling plans usually focus on one assignable cause. However, multiple assignable causes may occur especially for a multistage production system, and the resulting process shift may propagate downstream. This paper addresses the problem of finding the optimal sampling plan for an unreliable multistage production system subject to competing and propagating random quality shifts. In particular, a serial production system with two unreliable machines that produce a product at a fixed production rate is studied. It is assumed that both machines are subject to random quality shifts with increased nonconforming rates and can suddenly fail with increasing failure rates. A sampling plan is implemented at the end of the production line to determine whether the system has shifted or not. If a process shift is detected, a necessary maintenance action will be initiated. The optimal sample size, sampling interval, and acceptance threshold are determined by minimizing the long-run cost rate subject to the constraints on average time to signal a true alarm, effective production rate, and system availability. A numerical example on an automatic shot blasting and painting system is provided to illustrate the application of the proposed sampling plan and the effects of key parameters and system constraints on the optimal sampling plan. Moreover, the proposed model shows better performance for various cases than an alternative model that ignores shift propagation.

Keywords: Sampling plan, multistage production systems, competing and propagating random shifts

1. Introduction

Quality improvement is a major concern for the success of a manufacturing enterprise. To be competitive, companies often adopt different procedures to improve their production processes for better product quality. However, regardless of the advances in technology and automation, a manufacturing environment is always subject to variability and random shift that affect product quality. As a result, it is important to perform process monitoring so that necessary actions can be taken for maintenance and process adjustments when the product's quality drops below an acceptable level.

Product inspection is one of process monitoring methods to determine if a process has shifted or not. The out-of-control state is attributed to the presence of assignable cause(s) such as tool wear, temperature increase, and wrong setups. Specially, an assignable cause makes a process variable, such as the process mean, to deviate from its target, or causes an attribute, such as the proportion of nonconformity, to increase. In addition to process shift, the production system may fail and stop production. When a process shift or system failure is detected, maintenance actions are initiated. Maintenance could be perfect, imperfect, or minimal. In particular, perfect maintenance restores a production unit to its good-as-new condition, imperfect maintenance restores the unit to a condition between its good-as-new and bad-as-old states, and minimal repair makes the unit operational while keeping the unit in the same health condition as before.

Regarding inspection options, screening (100% inspection), no inspection, sampling plans by control charts (online sampling), acceptance sampling, and continuous sampling are the most widely used. In practice, an inspection policy is adopted according to the type of production and a specific goal. For instance, acceptance sampling is used for batch (lot) production to decide whether a batch should be accepted or not. Such inspection procedures can be employed in both single-stage and multistage systems. Specially, a multistage system is composed of multiple components, machines, processes, or stages required to make the final product (Shi and Zhou, 2009).

A sampling plan is either designed economically or economically-statistically. Economic designs aim at minimizing a cost function without focusing on statistical performance, while economic-statistical designs consider the performance of a process under some practical constraints. The usual performance metrics could be customer-centered such as the average outgoing quality (AOQ). Some measures are more producer-centered such as the average fraction inspected (AFI), process availability, and throughput. Other metrics, such as schedules' delays, are concerning both parties. Studies on these measures can be found in Bouslah et al. (2013), Cao and Subramaniam (2013), and Pandey et al. (2011). Existing sampling plans are often developed based on one assignable cause. Although a few studies consider cases with multiple assignable causes, it is often assumed that only one assignable cause can occur during a sampling cycle.

In this paper, we develop an economic-statistical sampling plan for a serial production system with two unreliable machines by considering the occurrences of more than one assignable cause. The term "stage" can be used in lieu of "machine" to refer to a process or a group of machines (processes). The sampling plan is modeled based on the competency and downstream propagation of process shifts. Sampling parameters are determined by minimizing the long-run cost rate subject to constraints on effective production rate, average time to signal a true alarm and system availability. It is assumed that sampling is performed only after the second stage. For example, in some systems, the synchronized handling of products from one stage to another does not allow any stoppage for inspection after the first stage. In other systems, products are processed sequentially or simultaneously by two different processes on the same machine making quality inspection impractical due to the machine's complex configuration.

Some industrial applications of such a system are as follows. In an automatic blasting and painting line, a fabricated steel unit is first blasted for rust removal and then fed into a painting chamber. Due to degradation, the disc turbines that provide blasting may still leave some rust on the unit's surface that causes poor paint adhesion. On the other hand, the spray nozzles in the painting chamber, if clogged, could cause bad paint coverage. The unit produced is nonconforming if one or both of the quality issues occur. An example of two processes being performed automatically on one machine is the production of purlins for steel structures. Galvanized sheets are fed continuously into a forming machine. Punching holes and bending edges are sequentially or simultaneously processed to produce a purlin. Due to the complex configuration of the machine, any quality imperfection cannot be observed until the whole process is complete. When the punching tips and/or the bending rollers become worn, the purlin is defective because holes, edges, or both are imprecisely made. Other examples in automotive painting and stamping lines are provided by Naebulharam and Zhang (2014). In some industries, inspection may be performed only after the final stage due to safety or economic reasons. For instance, small steel bars are first heated and then forged to produce small parts such as socket wrenches. Other examples are manufacturing of aluminum cans, automated bakery production, powder coating, automatic riveting for stamping parts, automatic assembling and wire bonding, and multi-material additive manufacturing of electronic devices. More applications of such systems are addressed by Liberopoulos et al. (2010).

The remainder of this paper is organized as follows: Section 2 reviews the related literature and illustrates the contributions. Section 3 describes the problem and the assumptions, and provides the notation used throughout this paper. A comprehensive modelling methodology is developed in Section 4. Section 5 provides the mathematical formulation for the optimal design of the proposed sampling plan. A numerical example and analyses are given in Section 6. Section 7 concludes this study and recommends several directions for future research.

2. Literature review and research contributions

2.1. Related work

In the context of single-stage production systems, Linderman et al. (2005) propose an economic-statistical cost model considering constraints on the average run lengths and three maintenance scenarios. Charongrattanasakul and Pongpullponsak (2011) extend this work by sampling with an exponentially weighted moving average (EWMA) chart with warning limits along with maintenance at the time of a false alarm. Mehrafrooz and Noorossana (2011) consider an additional maintenance scenario due to sudden machine failures. Pandey et al. (2011) use an $\bar{X}$ control chart to determine the sequence of batches produced on a single machine subject to scheduled preventive maintenance. Safaei et al. (2015) study sampling by an $\bar{X}$ control chart under uncertainty. Pasha et al. (2018) incorporate the Taguchi loss function in the design of $\bar{X}$ control chart with non-normal quality data. Abolmohammadi et al. (2019) develop an economical statistical design for variable parameters $\bar{X}$ control charts under different quality loss functions. It is worth pointing out that all these studies focus only on one assignable cause. However, this may not be realistic.

Indeed, multiple assignable causes from different sources, such as raw materials, human errors and tool wear, cannot be ignored. Yu and Hou (2006) develop an economic model for an $\bar{X}$ control chart with variable sampling intervals to monitor a process with multiple assignable causes. Yu et al. (2010) construct an economic-statistical model with constraints on type-I and type-II errors. The same constraints are used by Salmasnia et al. (2017). The effects of non-normal quality data on the design of $\bar{X}$ control chart with the presence of multiple assignable causes are investigated by Moghadam et al. (2018). Unlike these studies where only one assignable cause is permitted to occur during an inspection cycle, a case allowing the occurrences of multiple assignable causes during an inspection cycle is examined by Yang et al. (2010). An $\bar{X}$ control chart is designed, but the joint effect of two assignable causes is assumed to be the same. Xiang (2013) study the joint optimization of an $\bar{X}$ control chart and preventive maintenance for a deteriorating production system. The system is assumed to have multiple degraded states that correspond to different assignable causes, and an economic cost model for maintenance, operation, and inspection is provided.

Inspection procedures for multistage systems are diverse. Zantek et al. (2002) assume that the variation of a measurement at a stage depends on both the variation of process parameters (i.e., pressure, temperature, etc.) at the present stage and the variations of measurements taken at preceding stages. Their engineering model aims at identifying which quality and process variables are responsible for the variation at the final stage. Zhou et al. (2003) propose an engineering model for an automotive engine heads machining line. Without process variables, Lam et al. (2005) develop an engineering model for a four-stage machining process where the last stage has two streams (parallel machines), and each stage or stream is monitored by a separate $\bar{X}$ control chart. It is assumed that only one stage is out-of-control at any time and the probability that a stage is out-of-control is constant. The $\bar{X}$ control charts are only designed to alert out-of-control signals according to a desired average time to signal without addressing whether any adjustment on the process or any rework on defective products is carried out or not. Xiang and Tsung (2008) study statistical monitoring with EWMA control charts based on engineering models. The EWMA control chart is designed for a given in-control average run length to determine the out-of-control condition in a three-stage process where wrong fixturing causes the process to be out-of-control. An engineering model based on multivariate control charts to detect mean shifts with autocorrelated observations is proposed by Kim et al. (2017).

Inspection allocation is another focus related to multistage systems. Bai and Yun (1996) consider a serial three-stage circuit board manufacturing system with two inspection stations. Inspection locations and inspection level (number of components tested on a circuit board) are determined to minimize the expected total cost of rework, inspection, and defective boards delivered to customers. Rau and Chu (2005) study inspection allocation in a serial multistage system where inspection could be on product variables and attributes. Azadeh et al. (2015) study a batch production system where inspection allocation, inspection tolerances, and full inspection or acceptance sampling are determined. Types and locations of inspection are determined in a serial multistage system by the trade-off between production costs and customer satisfaction under uncertainty (Mohammadi et al. 2018).

The quality and quantity are the two main focuses of a multistage production system. Cao and Subramaniam (2013) investigate a serial multistage system where each stage is monitored by a continuous sampling plan (CSP). The CSP alternates between 100% and fractional inspections based on whether or not a consecutive number of conforming units are observed. Additional measures of work in process (WIP) and throughput rate are also considered. Kim and Gershwin (2005) study a two-machine system with one buffer using a Markov process. In their work, a machine is assumed to have three states: operating producing good parts, operating producing bad parts (quality failure state), and complete failure. The effects of quality failure, production rate, and buffer size on the system's yield and effective production rate are analyzed. Kim and Gershwin (2008) also analyze the performance of flow lines with quality and operational failures. Meerkov and Zhang (2010) investigate different cases for performance analysis of a serial production system with inspection stations and buffers under 100% inspection. Given the number of inspection stations and buffers capacities, the study shows the impact of inspection allocation on bottlenecks, blocked and starving machines, and effective production rate. Colledani and Tolio (2012) develop a Markovian model for a serial system subject to degradation. The critical state that separates the desired degradation states from the undesired states is determined by achieving gains in system's yield and effective production rate. It is worth pointing out that engineering models are analytical tools for identifying sources of variation for quality improvement. Usually, a strategy with 100% inspection of variables is adopted. On the other hand, in most of inspection allocation models, 100% inspection or acceptance sampling are used with the purposes of locating inspection and determining a testing strategy or inspection level. For both types of models, maintenance is rarely studied.

Liu et al. (2013) study a serial system consisting of two identical units monitored by an $\bar{X}$ control chart. The value of process shift is assumed to be a constant no matter one or both units are in the quality failure state, and an inspection cycle is renewed by one of four maintenance scenarios. The system's performance is evaluated via economic and economic-statistical models with constraints on type-I and type-II errors. Zhu et al. (2016) investigate a serial four-stage process where attributes sampling is carried out at each stage. In their work, only quality failures are considered, and the sampling parameters are found by minimizing the expected total cost of inspection, scrap, and repair with respect to constraints on the average number of produced products between two false alarms. Zhong and Ma (2017) propose a joint control chart for a two-stage dependent serial system where the first and second stages are monitored by an $\bar{X}$ and a residual control chart, respectively. Eight maintenance scenarios are investigated for cost minimization with constraints on the average run lengths. For more studies on part quality inspection in multistage production systems, readers are referred to a recent review by Rezaei-malek et al. (2019).

2.2. Contributions of this work

Clearly, the effects of quality failures, machine failures and maintenance actions on the product quality and the effective production rate of a multistage production system are worthy of investigation. Although a plenty of studies have been conducted on online sampling for single-stage production systems, only a few studies have been done on multistage systems. Specially, there is a lack of research on online sampling of attribute data for multistage systems. This study aims at developing an attribute sampling plan for a serial system of two unreliable machines for discrete production. Different from the work of Liu et al. (2013), this work considers two nonidentical machines and allows a quality shift to propagate downstream. Indeed, competing process shifts and downstream propagation are two forms of natural interactions in a multistage system. To the best of our knowledge, modeling sampling plans by attributes with competing shifts in a multistage system with unreliable machines have not been studied (Yang et al., 2010; Zhu et al., 2016) in the literature although such a study will have a wide variety of industry applications. In addition, this work develops a comprehensive economic-statistical model with closed-form formulations and establishes a compromise between quality and quantity performances. Unlike the studies by Yang et al. (2010), Liu et al. (2013) and Xiang (2013) that focus only on quality-related performance, we consider a constraint on system's availability to increase production, and a constraint on effective production rate to increase the fraction of good products. Moreover, a constraint on average time to signal is also included. This model represents a first step that can be extended for a production line with more than two unreliable machines, multiple assignable causes, and different levels of maintenance actions. The economic benefit of the proposed model over existing studies that do not consider shift propagation is illustrated in this work.

3. Problem description

A serial production system consisting of two unreliable machines that operate continuously to produce discrete units of a product is considered. Each unit of the product is first processed at machine 1 followed by machine 2. Each machine has the proportion of nonconforming (PON) of p_{0m}, $m \in \{1,2\}$ when it is in-control. Due to assignable causes, PON may increase to p_{1m} so that the machine enters its out-of-control state. Each machine is subject to two issues: quality shift when the PON increases from p_{0m} to p_{1m}, and sudden machine breakdown (failure). Failures are observed immediately, whereas quality shifts can be detected only by inspection.

To inspect the finished units, an attribute sampling plan is employed at the end of the production line (i.e., after machine 2) to assess the performance of the production process and to initiate necessary maintenance actions. An inspected unit is classified as either conforming or nonconforming, and if a half-finished unit is nonconforming upstream (after machine 1), it remains nonconforming downstream. The power of detecting a process shift depends on the parameter setting of the sampling plan. Clearly, sampling may generate two kinds of errors: type I error and type II error. Type-I error (false alarm) is generated when a process signals an alarm given that the process has not shifted yet. Type-II error is generated when the sampling plan fails to signal a true alarm when the process has already shifted. Determining which machine(s) has/have shifted cannot be done unless the system is shut down for close inspections of the two machines. Therefore, whenever there is a failure or a shift, both machines are stopped for maintenance. However, when machines are shut down because of a false alarm, no maintenance is carried out and production resumes.

It is assumed that the time to shift for machine m follows the exponential distribution with a rate of λ_m (see Liu et al., 2013 and Xiang, 2013), whereas time to failure is assumed to follow the two-parameter Weibull distribution with an increasing failure rate (see Pandey et al., 2011) . During operation, if a machine fails, minimal repair is performed, which makes the machine operational but does not reduce its failure rate after repair. If a shift is detected, both machines are restored to their good-as-new conditions with PON of p_{0m} and age 0, and a new inspection cycle begins. Restoration can be either corrective or preventive. Corrective restoration is performed on the machine that has the shift, whereas preventive restoration resets the age of the machine that has not shifted to zero.

Whenever a true alarm is signaled, it is clear that at least one machine has shifted. Clearly, the time to shift on each machine is random. The system is said to be out-of-control if a shift on any of the machines has occurred, and hence, the stochastic competency between shifts (which shift occurs first) determines what out-of-control state the system is currently in, as will be illustrated in Section 4.1. In this regard, the sampling plan is designed to detect such competing and propagating shifts. Specially, a propagating shift occurs if one machine has already shifted but that shift is not detected until another shift takes place on the other machine. In particular, the production system is classified as a multistage multistate system. The system at any sampling time can be in one of four states: one in-control state, and three out-of-control states. The system's PON (p_s) can be represented by a set

$$p_s = \{p_0, p_1, p_2, p_3\},$$

where $p_0 = \phi(p_{01}, p_{02})$ represents that the system is in the in-control state (i.e., both machines are in control) and $\phi(\cdot,\cdot)$ is a function of machines' PONs; $p_1, p_2,$ and p_3 represent that the system is out-of-control with $p_1 = \phi(p_{11}, p_{02})$ being that only machine 1 has shifted, $p_2 = \phi(p_{01}, p_{12})$ being that only machine 2 has shifted, and $p_3 = \phi(p_{11}, p_{12})$ being that both shifts have occurred. Note that for the system's probability of nonconforming, p_0 can evolve to either p_1 or p_2, and p_1 or p_2 can evolve to p_3. Basically, p_s can be determined by:

$$p_s = \phi(p_{f1}, p_{f2}) = 1 - \prod_{m=1}^{2}(1 - p_{fm}), \tag{1}$$

where $f = \{0$, machine is in-control; 1, machine is out-of-control$\}$.

To study the process with competing and propagating shifts, the sampling plan with one assignable cause proposed by Lorenzen and Vance (1986) is used as the baseline. The sampling plan is illustrated in Figure 1. A new inspection cycle starts with both machines being in good-as-new conditions. Inspection continues until a true alarm is signaled. Therefore, the inspection cycle length is defined as the time since the beginning of sampling until the two machines are restored correctively and/or preventively back to their good-as-new conditions after a true alarm. After each "h" time units (called the sampling interval), N units are sampled and inspected. If the number of nonconforming units in this sample exceeds an acceptance threshold r, the two machines are investigated to determine if the out-of-control signal is a false alarm or indeed a true alarm. All the sampled units found to be nonconforming are rejected without rework.

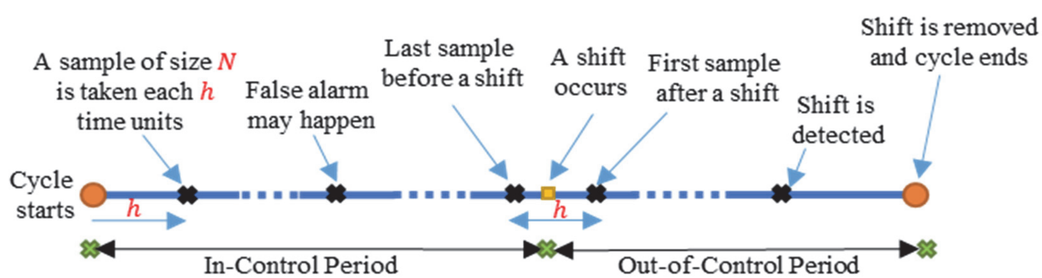

Figure 1. Sampling plan proposed by Lorenzen and Vance (1986).

By taking into account competing and propagating shifts, the sampling plan shown in Figure 1 is modified in Section 4. The objective is to design an attribute sampling plan considering stochastic competing and propagating shifts. An optimization model is developed to minimize the long-run cost rate and to find the optimal sampling parameters. The assumptions about system operation and the notation used in this paper are provided next.

Assumptions

- The raw materials are defect free (i.e., incoming quality is perfect). Note that if the incoming quality is not perfect, this effect can be folded into the first-stage in-control nonconforming probability.

- Quality shift and machine failure are independent. For example, in an automated painting line, as the ambient temperature decreases, paint becomes more viscous causing undesirable coat quality, but the increased viscosity of paint does not cause a complete machine failure.
- The occurrences of assignable causes that cause shifts on the two machines are independent, as the two machines perform different tasks, may run under different operating conditions, and are composed of different components. As will be explained in Section 6, the degradation of turbine discs causes a shift on the shot blasting machine, whereas the degradation of spraying nozzles causes another shift on the painting machine. Both shifts are independent as they occur on different machines without any linkage. Such assumptions about independent assignable causes (or shifts) have been made by others such as Yu et al. (2010), Xiang (2013), and Salmasnia et al. (2017).
- The production rates and reliability of the two machines are not significantly different.
- There are enough storage areas for the finished products and WIP so that the production will not be stopped because of lacking storage areas.
- The system is stopped during sampling, which prevents the process with a potential quality shift from running during sampling. This is reasonable if the loss due producing nonconforming units is high. Note that the sampling interval (i.e., h) is an important decision variable in this study.
- The two machines do not deteriorate or shift while being stopped.
- Maintenance requests can only be fulfilled in sequence. In other words, a machine can be maintained only after the current maintenance action is complete. This is reasonable when only one maintenance team is involved.

Notation

Decision variables

h	Sampling interval measured in hours.
N	Sample size
r	Acceptance threshold

Objective function

$LRCR$	Long-run cost rate measured in $/hour

Other variables, constants and indices

j	Index referring to the sample number at which an inspection cycle ends
i, k, q, w	Indices
m	Index for a machine, $m \in \{1,2\}$
G	Inspection cycle operational time excluding false alarms, minimal repairs, true alarm, and restoration times
S_m	Shift of machine m, $m \in \{1,2\}$
S_{12}	Propagating shift

λ_m	Shift rate of machine m, $m \in \{1,2\}$
T_m	Time to shift of machine m, exponentially distributed $T_m \sim \text{Exp}(\lambda_m)$, $m \in \{1,2\}$
τ_{S_m}	Time of occurrence of S_m since the last sampling
PON	Proportion of nonconforming
p_{fm}	PON of machine m, $m \in \{1,2\}$, $f = \{0$, machine m is in-control; 1, machine m is out-of-control$\}$
p_s	PON of the production system
$\phi(\cdot,\cdot)$	A function that represents p_s in terms of machines' PONs
d	Number of nonconforming units found in a sample of size N
α	Type-I error due to a false signal
T_{in}	Time process stays in the in-control state
T_{s_1}	Time the process is running with $p_s = p_1 = \emptyset(p_{11}, p_{02})$
T_{s_2}	Time the process is running with $p_s = p_2 = \emptyset(p_{01}, p_{12})$
$T_{s_{12}}$	Time the process is running with $p_s = p_2 = \emptyset(p_{11}, p_{12})$
β_{p_s}	Type-II error when $p_s \in \{p_1, p_2, p_3\}$
ARL_0	Average run length while the process is in-control
$ARL_{s_{12}}$	Average run length while the process is out-of-control with propagating shift
Q_{in}	Number of samples taken while the process is in-control
$Q_{p_1}(Q_{p_2})$	Number of samples taken while the process is operating with $p_s = p_1(p_2)$
$V_{in}(V_{out})$	Number of rejected units found during sampling in the in-control (out-of-control) period
RJU	Total number of rejected units during sampling
t_s	Average time of inspecting one unit of the product
$T_{FA}(T_{TA})$	Average time to search for a false (true) alarm on each machine
T_{MRm}	Average time to perform a minimal repair on machine m, $m \in \{1,2\}$
$CRT_m(PRT_m)$	Average corrective (preventive) restoration time on machine m, $m \in \{1,2\}$
S_t	Total time of sampling in an inspection cycle
TT_{FA}	Total time of searching for false alarms in one inspection cycle
TT_{TA}	Average total time of searching for a true alarm in an inspection cycle
MRT	Total time of minimal repairs in an inspection cycle
RT	Total restoration time in an inspection cycle
C_s	Average inspection cost per unit time
$C_{FA}(C_{TA})$	Average cost per unit time of searching for a false (true) alarm
C_{MR}	Average cost per unit time of performing a minimal repair
$C_{Cm}(C_{Pm})$	Average corrective (preventive) restoration cost per unit time for machine m, $m \in \{1,2\}$
C_{LP}	Average lost production cost per one unit of the product
C_{RJ}	Average cost of a rejected unit found during sampling
C_{NC}	Average cost of a nonconforming unit received by a consumer
S_c	Total cost of sampling in an inspection cycle
FA_c	Total cost of searching for false alarms in an inspection cycle
TA_c	Average total cost of searching for a true alarm in an inspection cycle
MR_c	Total cost of minimal repairs in an inspection cycle

$RC_{S_1}(RC_{S_2})$	Average restoration cost if an inspection cycle ends with $S_1(S_2)$
$RC_{S_{12}}$	Average restoration cost if an inspection cycle ends with S_{12}
RC	Total restoration cost in an inspection cycle
LP_c	Lost production cost in an inspection cycle
CRJ	Total cost of rejected units during sampling
CNC	Total cost of nonconforming units received by customers
$\theta_m(\gamma_m)$	Shape (scale) factor of Weibull distribution of machine m, $m \in \{1,2\}$, $\theta_m > 1$
g_m	Production rate of stage m
g_s	Production rate of the system, $\min_{m \in \{1,2\}} \{g_m\}$
$h_m(t)$	Failure rate of machine m, $m \in \{1,2\}$
$M_m(t)$	Expected number of failures of machine m, $m \in \{1,2\}$ in time interval $[0, t]$
MN_m	Number of minimal repairs on machine m, $m \in \{1,2\}$ in an inspection cycle
AV	System's availability
PR_{eff}	Effective production rate
ATS	Average time to signal
$CP(NCP)$	Number of conforming (nonconforming) products produced in one inspection cycle
TP	Total number of products produced in one inspection cycle
CC	Inspection cycle total cost
CT	Inspection cycle total time

4. Model development

4.1. Stochastic cases

Let G be the time at which the inspection cycle terminates due to detecting a shift. The random variable $G \in \{h, 2h, \cdots\cdots, \infty\}$ is the operational time that does not include the stoppage times of inspection, false alarms, minimal repairs, true alarms, and restorations, where the sampling interval h is the time between two successive inspections. Clearly, the shortest length of G is h. Since the production process has competing and propagating shifts, G can be derived based on the following three cases:

- Case I: Machine 2 shift (S_2) and machine 1 shift (S_1) occur in the same sampling interval, i.e., between $(i-1)h^{th}$ and ih^{th} sampling points as shown in Figure 2.
- Case II: S_2 is not detected before the occurrence of S_1 given that S_2 occurs between $(i-1)h^{th}$ and ih^{th} sampling points, and S_1 occurs after the ih^{th} sampling point as shown in Figure 3.
- Case III: S_2 is detected before the occurrence of S_1 as shown in Figure 4.

It is worth pointing out that the above cases also apply when S_1 occurs before S_2.

Case I. Let T_1 and T_2 be the times to shift of machines 1 and 2, respectively, and T_1 and T_2 follow the exponential distributions with rates λ_1 and λ_2, respectively. Moreover, let τ_{S_1} and τ_{S_2} be the times of occurrence of S_1 and S_2, respectively, since the most recent sampling. As shown in Figure 2, when $T_1 > T_2$, S_2 is missed because it is followed by S_1 before taking the next sample. Then, the production process starts to produce units with propagating shift at the time of occurrence of S_1.

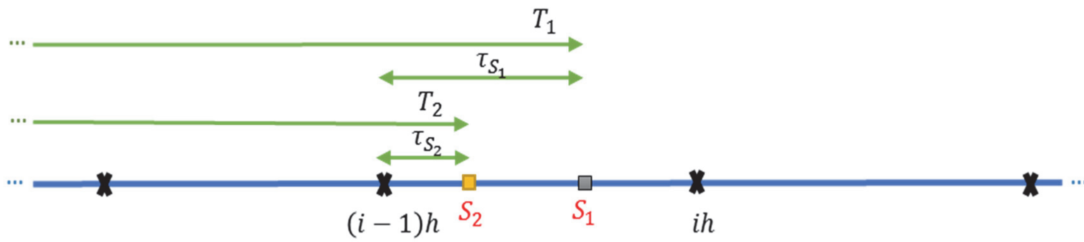

Figure 2. Case I, $T_1 > T_2$.

The probability that S_2 and S_1 happen in the same sampling interval given that $T_1 > T_2$ is

$$P\big((i-1)h \le T_2 \le T_1 < ih\big) = \int_{(i-1)h}^{ih} \int_{(i-1)h}^{t_1} \lambda_2 e^{-\lambda_2 t_2} \lambda_1 e^{-\lambda_1 t_1} dt_2\, dt_1$$

$$= e^{-\lambda_2 (i-1)h}\big(e^{-\lambda_1 (i-1)h} - e^{-\lambda_1 ih}\big) + \frac{\lambda_1}{\lambda_1+\lambda_2}\big(e^{-(\lambda_1+\lambda_2)ih} - e^{-(\lambda_1+\lambda_2)(i-1)h}\big).$$

Thus, the probability that $G = jh$ given that S_1 and S_2 happen between the $(i-1)^{th}$ and i^{th} sampling points and $T_1 > T_2$ is

$$P\big(G = jh, \text{Case I}_{T_1>T_2}\big) = \sum_{i=1}^{j} P\big((i-1)h \le T_2 \le T_1 < ih\big)\, \beta_{p_3}^{j-i}\big(1-\beta_{p_3}\big), \qquad j = 1, \cdots, \infty, \tag{2}$$

where β_{p_3} is the type II error resulting from that the system is producing units with $p_s = p_3 = p_{11} + p_{12} - p_{11}p_{12}$ according to equation 1. Let d be the number of nonconforming units in the sample, then the type II error $\beta_{p_s \in \{p_1,p_2,p_3\}}$ for $p_s \in \{p_1, p_2, p_3\}$ is given as

$$\beta_{p_s \in \{p_1,p_2,p_3\}} = \sum_{d=0}^{r} \binom{N}{d} p_s^d \,(1-p_s)^{N-d}. \tag{3}$$

For instance, in Case I and $T_1 > T_2$, $G = 2h$ if $0 \le T_2 \le T_1 < h$ and a shift is not detected until $j = 2$, or $h \le T_2 \le T_1 < 2h$ and a shift is detected at $j = 2$. Then, the probability that $G = 2h$ is

$$\left\{\big(1 - e^{-\lambda_1 h}\big) + \frac{\lambda_1}{\lambda_1+\lambda_2}\big(e^{-(\lambda_1+\lambda_2)h} - 1\big)\right\}\beta_{p_3}\big(1-\beta_{p_3}\big)$$
$$+\left\{e^{-\lambda_2 h}\big(e^{-\lambda_1 h} - e^{-\lambda_1 2h}\big) + \frac{\lambda_1}{\lambda_1+\lambda_2}\big(e^{-(\lambda_1+\lambda_2)2h} - e^{-(\lambda_1+\lambda_2)h}\big)\right\}\big(1-\beta_{p_3}\big).$$

The same procedure is followed for $T_2 > T_1$. Hence, $P\big((i-1)h \le T_1 \le T_2 < ih\big)$ and $P(G = jh, \text{Case I}_{T_2>T_1})$ can be expressed as follows, respectively:

$$P\big((i-1)h \le T_1 \le T_2 < ih\big) = \int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_2} \lambda_1 e^{-\lambda_1 t_1}\lambda_2 e^{-\lambda_2 t_2} dt_1\, dt_2$$

$$= e^{-\lambda_1(i-1)h}\big(e^{-\lambda_2(i-1)h} - e^{-\lambda_2 ih}\big) + \frac{\lambda_2}{\lambda_1+\lambda_2}\big(e^{-(\lambda_1+\lambda_2)ih} - e^{-(\lambda_1+\lambda_2)(i-1)h}\big),$$

$$P\big(G = jh, \text{Case I}_{T_2>T_1}\big) = \sum_{i=1}^{j} P\big((i-1)h \le T_1 \le T_2 < ih\big)\, \beta_{p_3}^{j-i}\big(1-\beta_{p_3}\big), \qquad j = 1, \dots, \infty. \tag{4}$$

Case II. As shown in Figure 3, S_1 occurs at least one sample after the occurrence of S_2. Due to the type II error, S_2 is always undetected until after the occurrence of S_1. The minimum value of G is $2h$ as a result that S_2 happens before taking the first sample (i.e., before time h) but is not detected, S_1 occurs afterwards, and the total shift is detected at time $2h$. If S_2 occurs in the sampling interval $[(i-1)h, ih]$, then S_1 could occur in any subsequent interval $[(i+k)h, (i+1+k)h]$ where $0 \le k \le j-i-1$ for any i, $1 \le i \le j-1$ and $j \ge 2$. Note that a true alarm is alerted at $j \ge i+1+k$, and hence, $k \le j-i-1$.

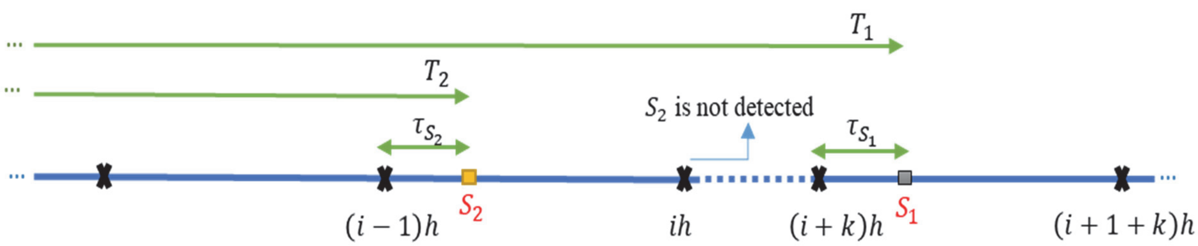

Figure 3. Case II, $T_1 > T_2$.

The probability that $G = jh$ in Case II and $T_1 > T_2$ is

$P(G = jh, \text{Case II}_{T_1>T_2}) =$

$$\sum_{i=1}^{j-1}\sum_{k=0}^{j-i-1} \big(e^{-\lambda_2(i-1)h} - e^{-\lambda_2 ih}\big)\big(e^{-\lambda_1(k+i)h} - e^{-\lambda_1(k+1+i)h}\big)\beta_{p_2}^{k+1}\beta_{p_3}^{j-i-k-1}\big(1-\beta_{p_3}\big), j = 2, \dots, \infty, \tag{5}$$

where β_{p_2} is the type II error (obtained by equation 3) that could result if the system is producing units with $p_s = p_2 = p_{01} + p_{12} - p_{01}p_{12}$. For instance, $P\big(G = h, \text{Case II}_{T_1>T_2}\big) = 0$, and $P\big(G = 2h, \text{Case II}_{T_1>T_2}\big) = \big(1 - e^{-\lambda_2 h}\big)\big(e^{-\lambda_1 h} - e^{-\lambda_1 2h}\big)\beta_{p_2}\big(1-\beta_{p_3}\big)$, and so on.

The same procedure can be followed for $T_2 > T_1$, and $P(G = jh, \text{Case II}_{T_2>T_1})$ is obtained as

$P(G = jh, \text{Case II}_{T_2>T_1}) =$

$$\sum_{i=1}^{j-1}\sum_{k=0}^{j-i-1} \big(e^{-\lambda_1(i-1)h} - e^{-\lambda_1 ih}\big)\big(e^{-\lambda_2(k+i)h} - e^{-\lambda_2(k+1+i)h}\big)\beta_{p_1}^{k+1}\beta_{p_3}^{j-i-k-1}\big(1-\beta_{p_3}\big), j = 2, \dots, \infty, \tag{6}$$

where β_{p_1} is the type II error (obtained by equation 3) that could result if the system is producing units with $p_s = p_1 = p_{11} + p_{02} - p_{11}p_{02}$.

Case III. In this case, as shown in Figure 4, S_2 is always detected at time $jh, j \geq i$, and before the occurrence of S_1. The probability that $G = jh$ given Case III and $T_1 > T_2$ can be expressed as

$$P\left(G = jh, \text{Case III}_{T_1>T_2}\right) = e^{-\lambda_1 jh} \sum_{i=1}^{j} \left(e^{-\lambda_2 (i-1)h} - e^{-\lambda_2 ih}\right) \beta_{p_2}^{j-i} \left(1 - \beta_{p_2}\right), \qquad j = 1, \dots, \infty. \tag{7}$$

For example, $P(G = h, \text{Case III}_{T_1>T_2}) = e^{-\lambda_1 h}\left(1 - e^{-\lambda_2 h}\right)\left(1 - \beta_{p_2}\right)$, and $P(G = 2h, \text{Case III}_{T_1>T_2}) = e^{-\lambda_1 2h}\left\{\left(1 - e^{-\lambda_2 h}\right)\beta_{p_2}\left(1 - \beta_{p_2}\right) + \left(e^{-\lambda_2 h} - e^{-\lambda_2 2h}\right)\left(1 - \beta_{p_2}\right)\right\}$, and so on.

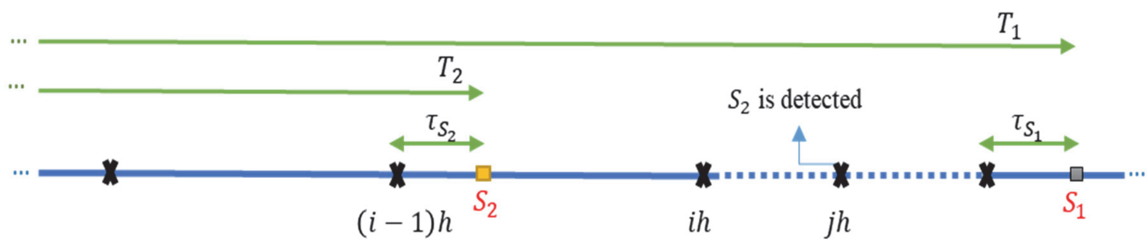

Figure 4. Case III, $T_1 > T_2$.

Similarly, when $T_2 > T_1$, $P(G = jh, \text{Case III}_{T_2>T_1})$ can be obtained as

$$P\left(G = jh, \text{Case III}_{T_2>T_1}\right) = e^{-\lambda_2 jh} \sum_{i=1}^{j} \left(e^{-\lambda_1 (i-1)h} - e^{-\lambda_1 ih}\right) \beta_{p_1}^{j-i} \left(1 - \beta_{p_1}\right), \qquad j = 1, \cdots, \infty. \tag{8}$$

Consequently, following the above cases, the expected value $E[G]$ can be given as

$$E[G] = A_1 + A_2 + A_3 + A_4 + A_5 + A_6, \tag{9}$$

where A_1 to A_6 are the weighted expected values of the cycle length given all cases. A_1 to A_6 are obtained as follows, respectively:

$$A_1 = \sum_{j=1}^{\infty} jh \cdot P\left(G = jh, \text{Case I}_{T_1>T_2}\right) = \frac{h(e^{(\lambda_1+\lambda_2)h} - \beta_{p_3})(\lambda_2 e^{\lambda_2 h}(e^{\lambda_1 h} - 1) - \lambda_1 (e^{\lambda_2 h} - 1))}{(\lambda_1+\lambda_2)(1-\beta_{p_3})\left(e^{(\lambda_1+\lambda_2)h} - 1\right)^2},$$

$$A_2 = \sum_{j=1}^{\infty} jh \cdot P\left(G = jh, \text{Case I}_{T_2>T_1}\right) = \frac{h(e^{(\lambda_1+\lambda_2)h} - \beta_{p_3})(\lambda_1 e^{\lambda_1 h}(e^{\lambda_2 h} - 1) - \lambda_2 (e^{\lambda_1 h} - 1))}{(\lambda_1+\lambda_2)(1-\beta_{p_3})\left(e^{(\lambda_1+\lambda_2)h} - 1\right)^2},$$

$$A_3 = \sum_{j=2}^{\infty} jh \cdot P\left(G = jh, \text{Case II}_{T_1>T_2}\right) = \frac{h\beta_{p_2}(e^{\lambda_1 h}-1)(e^{\lambda_2 h}-1)(e^{\lambda_1 h}+(\beta_{p_3}-2)e^{(2\lambda_1+\lambda_2)h}+\beta_{p_2}(e^{(\lambda_1+\lambda_2)h}-\beta_{p_3}))}{(\beta_{p_3}-1)(e^{(\lambda_1+\lambda_2)h}-1)^2(e^{\lambda_1 h}-\beta_{p_2})^2},$$

$$A_4 = \sum_{j=2}^{\infty} jh \cdot P\left(G = jh, \text{Case II}_{T_2>T_1}\right) = \frac{h\beta_{p_1}(e^{\lambda_1 h}-1)(e^{\lambda_2 h}-1)(e^{\lambda_2 h}+(\beta_{p_3}-2)e^{(\lambda_1+2\lambda_2)h}+\beta_{p_1}(e^{(\lambda_1+\lambda_2)h}-\beta_{p_3}))}{(\beta_{p_3}-1)(e^{(\lambda_1+\lambda_2)h}-1)^2(e^{\lambda_2 h}-\beta_{p_1})^2},$$

$$A_5 = \sum_{j=1}^{\infty} jh \cdot P(G = jh, \text{Case III}_{T_1>T_2}) = \frac{h(\beta_{p_2}-1)e^{\lambda_1 h}(e^{\lambda_2 h}-1)(\beta_{p_2}-e^{(2\lambda_1+\lambda_2)h})}{(e^{(\lambda_1+\lambda_2)h}-1)^2(e^{\lambda_1 h}-\beta_{p_2})^2},$$

$$A_6 = \sum_{j=1}^{\infty} jh \cdot P\big(G = jh, \text{Case III}_{T_2>T_1}\big) = \frac{h(\beta_{p_1}-1)e^{\lambda_2 h}(e^{\lambda_1 h}-1)(\beta_{p_1}-e^{(\lambda_1+2\lambda_2)h})}{(e^{(\lambda_1+\lambda_2)h}-1)^2(e^{\lambda_2 h}-\beta_{p_1})^2}.$$

4.2. Time and cost of sampling

The average number of samples taken during the inspection cycle equals to $E[G]/h$. Then, the expected time of sampling $E[S_t]$ can be expressed as

$$E[S_t] = \frac{t_s \cdot N \cdot E[G]}{h}, \tag{10}$$

where t_s is the average time of inspecting one unit of the product. Let C_s be the average cost per unit time of sampling, then the expected cost of sampling $E[S_c]$ is

$$E[S_c] = C_s\, E[S_t]. \tag{11}$$

4.3. Time and cost of false alarms

The process is out-of-control once any of the two shifts occurs. Consequently, the time period that the process is in-control T_{in} follows the exponential distribution with $T_{in} = \text{Min}(T_1, T_2) \sim \text{Exp}(\lambda_1 + \lambda_2)$. Therefore, the expected time that the process is in-control $E[T_{in}]$ is

$$E[T_{in}] = \frac{1}{\lambda_1 + \lambda_2}$$

Let Q_{in} be the number of samples taken when the system is in-control. Then, its expected value is

$$E[Q_{in}] = \sum_{i=0}^{\infty} i \cdot \left(e^{-(\lambda_1+\lambda_2)ih} - e^{-(\lambda_1+\lambda_2)(i+1)h}\right) = \frac{1}{e^{(\lambda_1+\lambda_2)h} - 1}.$$

As a result, the expected total time of false alarms $E[TT_{FA}]$ is given by

$$E[TT_{FA}] = 2\, T_{FA} \frac{E[Q_{in}]}{ARL_0}, \tag{12}$$

where T_{FA} is the average time for identifying a false alarm on each machine , ARL_0 is the average run length when the process is in-control (i.e., the average number of samples taken until a false alarm is alerted), and $E[Q_{in}]/ARL_0$ is the average number of false alarms in one cycle, in which ARL_0 is (Montgomery, 2009)

$$ARL_0 = \frac{1}{\alpha},$$

where the type-I error α is reported when $p_s = p_0 = p_{01} + p_{02} - p_{01}p_{02}$ and $d > r$, which is given by

$$\alpha = 1 - \sum_{d=0}^{r} \binom{N}{d} p_0^d\, (1 - p_0)^{N-d}.$$

The direct cost of false alarms is due to the effort taken for identifying false alarms and inspecting machines. Let C_{FA} be the average cost per unit time of searching for a false alarm. Then, the expected total cost of searching for false alarms can be expressed as

$$E[FA_c] = C_{FA}\, E[TT_{FA}]. \tag{13}$$

4.4. Time and cost of searching for a true alarm

Let C_{TA} be the average cost per unit time of searching for a true alarm, then the average total time TT_{TA} and cost TA_c of searching for a true alarm are given as follows, respectively:

$$TT_{TA} = 2\, T_{TA}, \tag{14}$$

$$TA_c = C_{TA}\ TT_{TA}. \tag{15}$$

4.5. Restoration time and cost

Restoration time is the time required for machine maintenance and shift removal(s). Since inspection ends with a shift, at least one of the two machines need corrective restoration. Three possible scenarios are described next.

- *Inspection cycle ends only with* S_1

For this scenario, machine 1 is correctively restored, and machine 2 is preventively restored. The probability that the inspection cycle ends with this scenario equals the probability that S_1 is detected before the occurrence of S_2. Let CRT_1 and PRT_2 be the average corrective restoration time of machine 1 and the average preventive restoration time of machine 2, respectively, and C_{C1} and C_{P2} be the average costs per unit time of corrective and preventive restorations on machines 1 and 2, respectively. Then, the average restoration cost of this scenario RC_{S_1} is

$$RC_{S_1} = C_{C1}\, CRT_1 + C_{P2}\, PRT_2.$$

- *Inspection cycle ends only with* S_2

In this scenario, machine 2 is correctively restored, and machine 1 is preventively restored. The probability that the inspection cycle ends in this scenario is the probability that S_2 is detected before the occurrence of S_1. Let PRT_1 and CRT_2 be the average preventive restoration time of machine 1 and the average corrective restoration time of machine 2, respectively, and C_{C2} and C_{P1} be the average costs per unit time of corrective and preventive restorations on machines 2 and 1, respectively. Then, the average restoration cost of this scenario RC_{S_2} is

$$RC_{S_2} = C_{P1}\, PRT_1 + C_{C2}\, CRT_2.$$

- *Inspection cycle ends with propagating shift* S_{12}

In this scenario, both machines have shifted, and corrective restorations are carried out on both machines. The average cost of restoration of this scenario $RC_{S_{12}}$ is given as

$$RC_{S_{12}} = C_{C1}\, CRT_1 + C_{C2}\, CRT_2.$$

Hence, the expected total restoration cost $E[RC]$ and time $E[RT]$ are given as follows, respectively:

$$E[RC] = RC_{S_1}\, B_6 + RC_{S_2}\, B_5 + RC_{S_{12}}\, B, \tag{16}$$

$$E[RT] = (CRT_1 + PRT_2)\, B_6 + (PRT_1 + CRT_2)\, B_5 + (CRT_1 + CRT_2)\, B, \tag{17}$$

where $B_1(B_2)$ is the probability of Case I given $T_1 > T_2(T_2 > T_1)$, $B_3(B_4)$ is the probability of Case II given $T_1 > T_2(T_2 > T_1)$, and $B_5(B_6)$ is the probability of Case III given $T_1 > T_2(T_2 > T_1)$. B, and B_1 to B_6 are given as follows, respectively:

$$B = B_1 + B_2 + B_3 + B_4,$$

$$B_1 = \sum_{j=1}^{\infty} P\big(G = jh, \text{Case I}_{T_1>T_2}\big) = \frac{\lambda_1\big(1-e^{\lambda_2 h}\big)+\lambda_2\big(e^{(\lambda_1+\lambda_2)h}-e^{\lambda_2 h}\big)}{(\lambda_1+\lambda_2)\big(e^{(\lambda_1+\lambda_2)h}-1\big)},$$

$$B_2 = \sum_{j=1}^{\infty} P\big(G = jh, \text{Case I}_{T_2>T_1}\big) = \frac{\lambda_2\big(1-e^{\lambda_1 h}\big)+\lambda_1\big(e^{(\lambda_1+\lambda_2)h}-e^{\lambda_1 h}\big)}{(\lambda_1+\lambda_2)\big(e^{(\lambda_1+\lambda_2)h}-1\big)},$$

$$B_3 = \sum_{j=2}^{\infty} P\big(G = jh, \text{Case II}_{T_1>T_2}\big) = \frac{\beta_{p_2}\big(e^{\lambda_1 h}-1\big)\big(e^{\lambda_2 h}-1\big)}{\big(e^{(\lambda_1+\lambda_2)h}-1\big)\big(e^{\lambda_1 h}-\beta_{p_2}\big)},$$

$$B_4 = \sum_{j=2}^{\infty} P\big(G = jh, \text{Case II}_{T_2>T_1}\big) = \frac{\beta_{p_1}\big(e^{\lambda_2 h}-1\big)\big(e^{\lambda_1 h}-1\big)}{\big(e^{(\lambda_1+\lambda_2)h}-1\big)\big(e^{\lambda_2 h}-\beta_{p_1}\big)},$$

$$B_5 = \sum_{j=1}^{\infty} P\big(G = jh, \text{Case III}_{T_1>T_2}\big) = \frac{e^{\lambda_1 h}\big(e^{\lambda_2 h}-1\big)\big(1-\beta_{p_2}\big)}{\big(e^{(\lambda_1+\lambda_2)h}-1\big)\big(e^{\lambda_1 h}-\beta_{p_2}\big)},$$

$$B_6 = \sum_{j=2}^{\infty} P\big(G = jh, \text{Case III}_{T_2>T_1}\big) = \frac{e^{\lambda_2 h}\big(e^{\lambda_1 h}-1\big)\big(1-\beta_{p_1}\big)}{\big(e^{(\lambda_1+\lambda_2)h}-1\big)\big(e^{\lambda_2 h}-\beta_{p_1}\big)}.$$

4.6. Time and cost of minimal repair

Minimal repair is performed each time a machine fails unless a shift is detected. By nature, minimal repair does not change the failure rate of a failed machine. The failure rate $h_m(t)$ of machine m is given as

$$h_m(t) = \frac{\theta_m}{\gamma_m}\left(\frac{t}{\gamma_m}\right)^{\theta_m - 1},$$

where $\theta_m > 1$ and γ_m are the corresponding shape and scale parameters of the Weibull distribution, respectively. Then, the expected number of failures (i.e., minimal repairs) $M_m(t)$ of machine m during the interval $[0, t]$ can be obtained as

$$M_m(t) = \int_0^t h_m(u)du = \left(\frac{t}{\gamma_m}\right)^{\theta_m}.$$

Since machines do not age during downtime, the expected number of minimal repairs on machine m in each inspection cycle $E[MN_m]$ can be expressed as

$$E[MN_m] = \sum_{j=1}^{\infty}\left(\frac{jh}{\gamma_m}\right)^{\theta_m} P(G = jh), \tag{18}$$

where

$$P(G = jh) = P\big(G = jh, \text{Case I}_{T_1>T_2}\big) + P\big(G = jh, \text{Case I}_{T_2>T_1}\big) + P\big(G = jh, \text{Case II}_{T_1>T_2}\big) +$$

$$P(G = jh, \text{Case II}_{T_2>T_1}) + P(G = jh, \text{Case III}_{T_1>T_2}) + P(G = jh, \text{Case III}_{T_2>T_1}).$$

Since the purpose of minimal repair is to make a failed machine operational again with minimal resources, the PON of the system will be the same as that right before the failure. Let T_{MRm} and $C_{MRm}, m \in \{1,2\}$ be the average time and cost per unit time to perform a minimal repair on machine m, respectively. Then the expected total time $E[MRT]$ and the expected total cost of performing minimal repairs $E[MR_c]$ are given as follows, respectively:

$$E[MRT] = T_{MR1}E[MN_1] + T_{MR2}E[MN_2], \tag{19}$$

$$E[MR_c] = C_{MR1}T_{MR1}E[MN_1] + C_{MR2}T_{MR2}E[MN_2]. \tag{20}$$

4.7. Cost of lost production

The time due to stoppages for searching for false alarms and true alarms, sampling, minimal repairs, and restoration causes loss in production. Let C_{LP} be the average cost of lost production per one unit of the product, then the expected cost of lost production $E[LP_c]$ can be expressed as

$$E[LP_c] = C_{LP}g_s\{E[TT_{FA}] + TT_{TA} + E[S_t] + E[MRT] + E[RT]\}, \tag{21}$$

where g_s is the system's production rate given as $g_s = \min_{m\in\{1,2\}}\{g_m\}$ where g_m is the production rate of machine m.

4.8. Cost of units rejected in all samples

Any nonconforming unit found in a sample is rejected without replacement, and the production process at each sampling time should be in one of the following states: in-control state and three out-of-control states. To find the cost of rejected units in all samples, we first define the following quantities:

$$a_{p_s} = \sum_{d=r+1}^{N} d\binom{N}{d} p_s^d\,(1-p_s)^{N-d}, p_s \in \{p_0, p_1, p_2, p_3\},$$

$$b_{p_s} = \sum_{d=0}^{r} d\binom{N}{d} p_s^d\,(1-p_s)^{N-d}, p_s \in \{p_0, p_1, p_2, p_3\},$$

where a_{p_s} represents the expected number of nonconforming units found in a sample if a false or a true alarm is alerted ,whereas b_{p_s} refers to the expected number of nonconforming units found in a sample taken if no alarm is alerted. For instance, $a_{p_1}$is the expected number of nonconforming units found in the last sample that alerts the true alarm when the process is operating with S_1 ,whereas b_{p_0} is the expected number of nonconforming units found in a sample taken while the process is in control and no false alarm is alerted.

Any sample taken in the in-control period may indicate no alarm or false alarm, and the expected number of samples with false alarms equals to the expected number of false alarms. Then, the expected number of rejected units found during inspection when the process is in-control $E[V_{in}]$ is

$$E[V_{in}] = \alpha E[Q_{in}]a_{p_0} + (1-\alpha)E[Q_{in}]b_{p_0}.$$

The expected total number of rejected units during inspection $E[V]$ is given as

$$E[V] = E[V_{in}] + E[V_{out}], \tag{22}$$

where $E[V_{out}]$ is the expected total number of rejected units found in the out-of-control state. The derivation of $E[V_{out}]$ is provided in the Appendix. Let C_{RJ} be the average cost of a rejected unit, then the expected cost of rejected units $E[CRJ]$ is

$$E[CRJ] = C_{RJ}E[V]. \tag{23}$$

4.9. Cost of nonconforming units delivered to customers

A nonconforming unit found by a customer may cost more than a nonconforming unit found during the inspection. Let C_{NC} be the average cost of a nonconforming unit received by a customer, then the expected cost of nonconforming units received by customers $E[CNC]$ is given by

$$E[CNC] = C_{NC}\{g_s(p_0E[T_{in}] + p_1E[T_{s_1}] + p_2E[T_{s_2}] + p_3E[T_{s_{12}}]) - E[V]\}, \tag{24}$$

where $E[T_{s_1}]$, $E[T_{s_2}]$, and $E[T_{s_{12}}]$ are the expected values of times that the process could operate with S_1, S_2, and S_{12}, respectively. The details of these terms are given in Section 5.

4.10. Expected total cycle cost and time

Based on the above calculations, the expected total cycle cost $E[CC]$ and the expected total cycle time $E[CT]$ can be obtained as follows, respectively:

$$E[CC] = E[S_c] + E[FA_c] + TA_c + E[RC] + E[MR_c] + E[LP_c] + E[CRJ] + E[CNC], \tag{25}$$

$$E[CT] = E[G] + E[S_t] + E[TT_{FA}] + TT_{TA} + E[RT] + E[MRT]. \tag{26}$$

5. Optimal design of the sampling plan

The optimal sampling parameters are determined by minimizing the long-run cost rate $LRCR = E[CC]/E[CT]$, which is the ratio between the expected total cycle cost and the expected total cycle time. The mathematical formulation of the problem is given by

$$\min_{N,r,h} \quad LRCR = \frac{E[CC]}{E[CT]} \tag{27}$$

Subject to

$$AV \geq A \tag{27.1}$$

$$PR_{eff} \geq W \tag{27.2}$$

$$ATS \leq L \tag{27.3}$$

$$N \leq (h - u_l)g_s, \quad l \in \{1,4,5,6\} \tag{27.4}$$

$$N > r \tag{27.5}$$

$$N, r \in \mathrm{Z}+, \quad h > 0. \tag{27.6}$$

The formulation belongs to a Mixed Integer Nonlinear Programming (MINLP) problem. Equation (27) states that $LRCR$ is minimized with respect to the three decision variables $N, r,$ and h. Equations (27.1) - (27.3) specify three performance constraints. In equation (27.1), the system availability AV must be greater than or equal to a predefined threshold A to ensure the expected total number of units produced in one cycle. However, with increased availability, both the expected numbers of conforming and nonconforming units increase. Since the latter is undesirable, equation (27.2) imposes another constraint on the effective production rate PR_{eff} to ensure the fraction of expected number of conforming units produced is above a certain level W. Moreover, equation (27.3) is used to ensure the speed of detecting process shifts in terms of the average time to signal ATS. ATS is defined as the average time taken to alert a true alarm since the occurrence of a shift. In practice, ATS could be short to avoid excess losses when producing products in the out-of-control state (i.e., ATS should be less than or equal to a threshold L). Inspection at each sampling time is carried out from the last unit produced, and a group of constraints given by equation (27.4) is provided to ensure that units are sampled from only one population (i.e., with the same p_s). These constraints also guarantee that N is always less than the number of units produced between two inspections. Note that because $u_1 > u_2$ when $T_1 > T_2$, we have $h - u_1 < h - u_2$. Moreover, because $u_4 > u_3$ when $T_2 > T_1$, we have $h - u_4 < h - u_3$ ($u_1$to u_6 are defined below). Therefore, the constraints corresponding to $l \in \{2,3\}$ are redundant. Lastly, the decision variables r and $N(> r)$ are nonnegative integers, and h is a positive continuous variable as specified in equations (27.5) and (27.6), respectively.

Since the three performance measures are essential to the operation of this system, they will be elaborated next.

System's availability

The system's availability AV is defined as:

$$AV = \frac{E[G]}{E[CT]}, \tag{28}$$

which is the ratio between the expected operational time in a cycle and the expected total cycle length.

Effective production rate

The effective production rate PR_{eff} is the proportion of the expected numbers of conforming units produced $E[CP]$ in the inspection cycle. PR_{eff} can be obtained as

$$PR_{eff} = \frac{E[CP]}{E[TP]} = 1 - \frac{E[NCP]}{E[TP]},$$

where $E[TP]$ and $E[NCP]$ are the expected total number and the expected number of nonconforming units produced in one cycle, respectively. $E[NCP]$ is the sum of the number of nonconforming units produced in the in-control state and the other three out-of-control states. Since each state has a different p_s, $E[NCP]$ and $E[TP]$ are given as follows, respectively:

$$E[NCP] = g_s\{p_0E[T_{in}] + p_1E[T_{s_1}] + p_2E[T_{s_2}] + p_3E[T_{s_{12}}]\},$$

$$E[TP] = g_sE[G].$$

Therefore, PR_{eff} is

$$PR_{eff} = 1 - \frac{\{p_0E[T_{in}] + p_1E[T_{s_1}] + p_2E[T_{s_2}] + p_3E[T_{s_{12}}]\}}{E[G]}, \quad (29)$$

where

$$E[T_{s_1}] = \{u_4 - u_3\}C_2 + C_4 + C_6,$$

$$E[T_{s_2}] = \{u_1 - u_2\}C_1 + C_3 + C_5,$$

$$E[T_{s_{12}}] = \left(hARL_{s_{12}} - u_1\right)C_1 + \left(hARL_{s_{12}} - u_4\right)C_2 + \left(hARL_{s_{12}} - u_5\right)C_7 + \left(hARL_{s_{12}} - u_6\right)C_8,$$

where $u_1(u_3)$ is the conditional expectation of τ_{S_1} given Case I, $T_1 > T_2(T_2 > T_1)$,whereas $u_2(u_4)$ is the conditional expectation of τ_{S_2} given Case I, $T_1 > T_2(T_2 > T_1)$, $u_5(u_6)$ is the conditional expectation of $\tau_{S_1}(\tau_{S_2})$ given Case II/III, $C_1(C_2)$ are the corresponding probabilities of Case I, $T_1 > T_2(T_2 > T_1)$, $C_3 = E\left[T_{s_2}, \text{Case II}_{T_1>T_2}\right]$, $C_4 = E\left[T_{s_1}, \text{Case II}_{T_2>T_1}\right]$, $C_5 = E\left[T_{s_2}, \text{Case III}_{T_1>T_2}\right]$, $C_6 = E\left[T_{s_1}, \text{Case II}_{T_2>T_1}\right]$, and $C_7(C_8)$ is the probability that the time needed is $hARL_{s_{12}} - u_5$ ($hARL_{s_{12}} - u_6$) to alert a true alarm since the occurrence of a shift given Case II, $T_1 > T_2(T_2 > T_1)$. The derivations of $E[T_{s_1}]$, $E[T_{s_2}]$, $E[T_{s_{12}}]$, u_1 to u_6, and C_1 to C_8 are given in the Appendix.

Average time to signal

As defined earlier, ATS is the average time taken until the sampling plan is successful to alert a true alarm since the occurrence of a shift. However, the process could run with two shifts (propagating shift), and hence, the exact definition of ATS will be the average time taken to alert a true alarm since the occurrence of the earlier shift. In Case I, as shown in Figure 2, S_1 or S_2 occurs first, and then, it propagates and becomes S_{12} until it is detected. The average number of samples taken to alert a true alarm is $ARL_{s_{12}}$, and hence, $ATS|$Case I is

$$ATS|\text{Case I} = \begin{cases} hARL_{s_{12}} - u_2, & T_1 > T_2 \\ hARL_{s_{12}} - u_3, & T_2 > T_1. \end{cases}$$

As shown in Figure 3 ($T_1 > T_2$), S_2 occurs τ_{S_2} time units since time $(i-1)h$. Therefore, $qh - u_6 + u_5$ is the elapsed time between the occurrences of S_2 and S_1. At the time of the occurrence of S_1 , the process starts operating with S_{12} until true detection, i.e., $hARL_{s_{12}} - u_5$ units time needed to alert a true alarm.

Summing up these times, $h(q + ARL_{s_{12}}) - u_6$ is the ATS since the occurrence of S_2. The same applies when $T_2 > T_1$, but with $h\left(q + ARL_{s_{12}}\right) - u_5$, and therefore, $ATS|$Case II is given as

$$ATS|\text{Case II} = \begin{cases} h(q + ARL_{s_{12}}) - u_6, & T_1 > T_2\,, q = \{1, \cdots, \infty\} \\ h\left(q + ARL_{s_{12}}\right) - u_5, & T_2 > T_1\,, q = \{1, \cdots, \infty\}, \end{cases}$$

where q refers to the number of samples taken between the occurrence times of the two shifts.

For Case III, as shown in Figure 4, there is no S_{12}. Therefore, $ATS|$Case III is

$$ATS|\text{Case III} = \begin{cases} wh - u_6, & T_1 > T_2\;\,, w = \{1, \cdots, \infty\} \\ wh - u_5, & T_2 > T_1\;\,, w = \{1, \cdots, \infty\}, \end{cases}$$

where w represents the number of samples that process undergoes with $S_2(S_1)$ until a successful detection. Note that $ATS\big|\text{Case III}_{T_1>T_2}$ and $ATS\big|\text{Case III}_{T_2>T_1}$ equal to the conditional expectations of T_{s_2} and T_{s_1}, respectively, given Case III as shown in the Appendix. Therefore C_5 and C_6 are used in the equation below.

Considering all cases, ATS is given by

$$ATS = (ATS|\text{Case I})\; C_1 + (ATS|\text{Case I})\; C_2 + D_1 + D_2 + C_5 + C_6, \tag{30}$$

where

$$D_1 = \sum_{q=1}^{\infty}\sum_{i=1}^{\infty} \left(ATS\big|\text{Case II}_{T_1>T_2}\right)\left(e^{-\lambda_2(i-1)h} - e^{-\lambda_2 ih}\right)\left(e^{-\lambda_1(i+q-1)h} - e^{-\lambda_1(i+q)h}\right)\beta_2^q =$$
$$\frac{\beta_2\left(h\left(ARL_{s_{12}}+1\right)-u_6\right)\left(e^{\lambda_1 h}-e^{2\lambda_1 h}-e^{(\lambda_1+\lambda_2)h}+e^{(2\lambda_1+\lambda_2)h}\right)+\beta_2^2\left(hARL_{s_{12}}-u_6\right)\left(e^{\lambda_1 h}+e^{\lambda_2 h}-e^{(\lambda_1+\lambda_2)h}-1\right)}{\left(e^{(\lambda_1+\lambda_2)h}-1\right)\left(e^{\lambda_1 h}-\beta_2\right)^2},$$

$$D_2 = \sum_{q=1}^{\infty}\sum_{i=1}^{\infty} \left(ATS\big|\text{Case II}_{T_2>T_1}\right)\left(e^{-\lambda_1(i-1)h} - e^{-\lambda_1 ih}\right)\left(e^{-\lambda_2(i+q-1)h} - e^{-\lambda_2(i+q)h}\right)\beta_1^q =$$
$$\frac{\beta_1\left(h\left(ARL_{s_{12}}+1\right)-u_5\right)\left(e^{\lambda_2 h}-e^{2\lambda_2 h}-e^{(\lambda_1+\lambda_2)h}+e^{(\lambda_1+2\lambda_2)h}\right)+\beta_1^2\left(hARL_{s_{12}}-u_5\right)\left(e^{\lambda_1 h}+e^{\lambda_2 h}-e^{(\lambda_1+\lambda_2)h}-1\right)}{\left(e^{(\lambda_1+\lambda_2)h}-1\right)\left(e^{\lambda_2 h}-\beta_1\right)^2}.$$

6. Numerical example and sensitivity analysis

We consider an automatic shot blasting and painting system as shown in Figure 5. Small fabricated steel parts such as cleats or rails are first loaded into the conveyor (or hanged on a monorail) and fed into the shot blasting chamber to remove rust from the surface of each part and texturizes it for better paint adhesion. Afterwards, parts are moved to the painting chamber for coating. Both blasting and painting are performed in closed environments. In the blasting machine, turbine disks that blow shot blasting balls on part surface are subject to degradation. Degradation of those disks reduces the amount of balls that hit the surface, so that possible rust could be left on the part's surface. On the other hand, the nozzles of spray guns in the painting chamber may be clogged so that they cannot uniformly spray paint and may dip some frozen paint particles on the part's surface. Indeed, painting on a rusty surface and dipping frozen paint particles cause a rough paint appearance. At the end of the line, a sampling plan by attributes explained previously is employed for inspecting the painted products. The deteriorated turbine disks and spray guns are considered as the sources of assignable causes, but they do not cause machines to breakdown. Instead, machine failures can be caused by other reasons such as overheating and power outage.

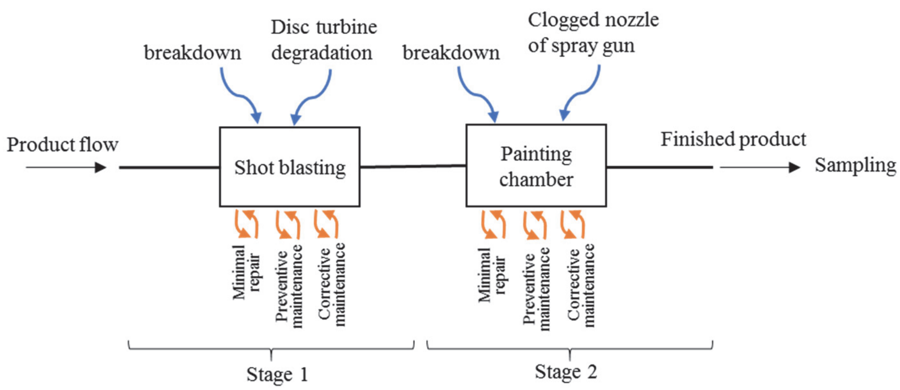

Figure 5. Automatic production line of shot blasting and painting.

Tables 1-3 show the parameters of shifts, failures, production rate, costs, time elements, and bounds of different constraints. T_{FA} is chosen to be greater than T_{TA}, as it is often easier to detect a shift when a process actually has shifted, whereas more time may be spent to verify that there is no shift in case of a false alarm. C_{FA} and C_{TA} are assumed to be equal as the same tooling and practices are required. The time and cost of maintenance increase as the degree of a maintenance action increases. Specially, corrective restoration may include replacing some components (e.g., turbine disk, spray gun, filter, nozzle) and thus require more tooling than other types of maintenance. However, a minimal repair needs the minimum resources to make the failed machine operational again. Therefore, we have $C_{cm} > C_{pm} > C_{MR}$ and $CRT_m > PRT_m > T_{MR}$. Moreover, since C_{NC} may include indirect costs such as claims and the company's goodwill, it is assumed that C_{NC} is greater than C_{LP} and C_{RJ}. The values of $\lambda_1(\lambda_2)$ shown in Tables 1 and 5 are chosen according to Zhong and Ma (2017), Mehrafrooz and Noorossana (2011), and Yang et al. (2010) where $0.001 \leq \lambda \leq 0.15$, whereas the values of $p_{01}(p_{02})$ and $p_{11}(p_{12})$ shown in Tables 1 and 4 are chosen with respect to the values used by Zhu et al. (2016) where $0.02 \leq p_0 \leq 0.04$ and $0.08 \leq p_1 \leq 0.12$.

Table 1. Shift and failure parameters, and production rate.

p_{01}	p_{11}	p_{02}	p_{12}	λ_1	λ_2	θ_1	θ_2	γ_1	γ_2	g_1, g_2
0.03	0.10	0.05	0.10	0.01 hr^{-1}	0.03 hr^{-1}	1.5	2.0	10 hr	10 hr	100,100 units/hr

Table 2. Cost parameters.

C_s	C_{c1}	C_{p1}	C_{c2}	C_{p2}	C_{MR1}	C_{MR2}	C_{FA}	C_{TA}	C_{LP}	C_{RJ}	C_{NC}
100 \$/hr	1200 \$/hr	600 \$/hr	1200 \$/hr	600 \$/hr	150 \$/hr	150 \$/hr	200 \$/hr	200 \$/hr	3.00 \$/unit	3.00 \$/unit	4.50 \$/unit

Table 3. Parameters of key time elements and bounds of constraints.

t_s	CRT_1	PRT_1	CRT_2	PRT_2	T_{MR1}	T_{MR2}	T_{FA}	T_{TA}	L	A	W
0.5 min/unit	50 min	25 min	50 min	25 min	15 min	15 min	15 min	7.5 min	3.00 hr	0.800	0.900

The MINLP problem given in Section 5 is mathematically complex since it has continuous and discrete decision variables and a discontinuous solution space. Moreover, the complex expressions involving discrete decision variables make the problem more complex. As a result, it is difficult to solve the optimization problem analytically or by an exact solution method. Instead, metaheuristics like Genetic Algorithm (GA) can be used. GA searches in parallel from a population of points so it can effectively explore many different solutions at the same time. When a certain solution turns out to be nonoptimal, GA discards it and proceeds with other more likely candidates. Therefore, GA does not tend to be easily trapped by local optima (Ahmed et al. 2014). In the literature, similar sampling plan problems have been solved using GA (e.g., Safaei et al., 2015; Abolmohammadi et al., 2019). Sultana et al. (2014) use both GA and Simulated Annealing (SA) in the economic design of $\bar{X}$ control chart, and the results show that GA provides solutions similar to SA but with less time. Moreover, GA is found superior (in terms of the quality solution obtained and the processing time) to SA, Particle Swarm Optimization (PSO), and Differential Evolution for the optimal design of multivariate EWMA (Malaki et al. 2011).

Due to the advantages of GA in solving such MINLP problems, especially those on sampling plans, GA in MATLAB R2019b is used in this work. In this study, the population size is twenty as only three decision variables are to be determined. The integer GA solver in MATLAB overrides settings supplied for creation, crossover, and mutation functions. Instead, GA uses special creation, crossover, and mutation functions (MATLAB & Simulink, 2019). To make the search process more efficient, strict constraint and function tolerance are used (set to default values, i.e., 1×10^{-3} and 1×10^{-6}, respectively). Moreover, the UseParallel option is used to compute the fitness value and the feasibility of nonlinear constraints in parallel to speed up the computation. The search process is stopped if any of the following criteria is met:

- The maximum number of generations (iterations) is reached. Here, the default number is used (i.e., 100 × number of decision variables).
- The average change in the penalty fitness value is less than the function tolerance over stall generations where the maximum stall generations is 50.
- Time limit is reached. Here, the default setting is used (i.e., infinity).

- There is no improvement in the objective function during an interval of time called stall time limit. Here, the default setting of the stall time limit is used (i.e., infinity).

The optimal solution is $LRCR^* = \$141.61$/hr, $r^* = 1$, $N^* = 5$, and $h^* = 0.428$ hrs. The optimization problem is solved many times with an average computational time of 133 seconds. To illustrate the economic benefits and the proper use of the proposed sampling plan in practice, an alternative design that allows only one assignable cause to occur in an inspection cycle is compared. Specially, the two designs are defined as follows:

- Model 1 (proposed in this paper) allows two assignable causes to occur in an inspection cycle.
- Model 2 considers that only one assignable cause can occur during an inspection cycle without considering shift propagation (e.g., Yu et al., 2010; Salmasnia et al., 2017). It is worth pointing out that Model 2 is similar to Case III in Model 1.

$LRCR_{M1}$ and $LRCR_{M2}$ are used as the objective functions of the two models, and their performance measures are investigated over a wide range of parameter settings. Moreover, the influence of the required ATS and the marginal effects of decision variables are also examined. The analysis is explained next.

Effect of PON(s) on models' performances. Collecting large data might be needed to estimate PON(s) parameters, and they depend on the machine's condition. To cope up with the uncertainty that could arise from imprecise estimation, the impact of those parameters on the performances of the two models is shown in Table 4. The parameters are changed by different percentages of the original setup (see Table 1). Since Model 2 allows only one shift to occur, as PON(s) are changed by $\geq +30\%$, more samples are taken, and the number of false alarms increases to alert an earlier true alarm. This increases the costs of false alarms, lost production and sampling, and reduces the cycle time. Hence, $LRCR_{M2} > LRCR_{M1}$when PON(s) are changed more than $+30\%$ where $PR_{eff} \leq 0.900$. This justifies why Model 2 has a larger (or equal) N compared to Model 1. Although the costs of Model 1 increase when PON(s) are changed less than +30%, this increase is absorbed by a longer cycle time making $LRCR_{M1} \approx LRCR_{M2}$. One can see that on average, $LRCR_{M2}$ is only 0.41% less than $LRCR_{M1}$ in the range from -50% to $+20\%$, whereas $LRCR_{M1}$ is 7.4% less than $LRCR_{M2}$ in the range from $+30\%$ to $+100\%$. This means that for the full range, Model 1 can be used.

Table 4. Effect of PONs on the optimal solutions of the two models.

	PON				Model 1				Model 2				
	p_{01}	p_{11}	p_{02}	p_{12}	r	N	h	$LRCR_{M1}$	r	N	h	$LRCR_{M2}$	PR_{eff}
−50%	0.015	0.05	0.025	0.05	0	2	0.417	122.52	0	2	0.450	121.03*	0.900
−40%	0.018	0.06	0.03	0.06	0	2	0.502	125.52	0	2	0.542	124.54*	0.900
−30%	0.021	0.07	0.035	0.07	0	2	0.588	129.22	0	2	0.634	128.73*	0.900
−20%	0.024	0.08	0.04	0.08	0	2	0.674	133.37*	0	2	0.727	133.37*	0.900

−10%	0.027	0.09	0.045	0.09	0	2	0.761	137.80*	0	2	0.820	138.18	0.900
0%	0.03	0.1	0.05	0.1	1	5	0.428	141.61	1	5	0.457	141.21*	0.900
+10%	0.033	0.11	0.055	0.11	1	5	0.507	143.00*	1	5	0.542	143.01	0.900
+20%	0.036	0.12	0.06	0.12	1	4	0.383	144.56	1	4	0.410	144.51*	0.900
+30%	0.039	0.13	0.065	0.13	1	4	0.442	146.18*	1	4	0.378	152.88	0.891
+40%	0.042	0.14	0.07	0.14	1	4	0.506	148.49*	1	5	0.534	161.63	0.884
+50%	0.045	0.15	0.075	0.15	1	4	0.573	151.34*	1	5	0.628	163.13	0.875
+60%	0.048	0.16	0.08	0.16	1	4	0.634	154.84*	1	4	0.405	168.12	0.868
+70%	0.051	0.17	0.085	0.17	1	4	0.704	158.47*	1	5	0.679	172.33	0.860
+80%	0.054	0.18	0.09	0.18	1	3	0.388	160.25*	1	4	0.443	176.24	0.853
+90%	0.057	0.19	0.095	0.19	1	3	0.434	162.42*	1	4	0.498	177.83	0.845
+100%	0.06	0.2	0.1	0.2	1	3	0.483	164.91*	1	3	0.308	180.27	0.837

Effect of quality shift parameters on models' performances. Parameters, λ_1 and λ_2 are related to the process that are difficult to estimate. These parameters are changed within wider ranges as shown in Table 5. High λ_1 and λ_2 increases the probability that shifts occur earlier, and hence, the probability of having a propagating shift increases. The costs of restoration and lost production increase since machines are highly likely to need corrective maintenance. Although the total cost increases more in Model 1, the increase is absorbed by a longer cycle time. This makes Model 1 more economical than Model 2 when λ_1 and λ_2 are high (i.e., $0.05 \leq \lambda_1 \leq 0.08$ and $0.07 \leq \lambda_2 \leq 0.1$) where $0.724 \leq A \leq 0.767$. For the medium ranges (i.e., $0.02 \leq \lambda_1 \leq 0.045$ and $0.04 \leq \lambda_2 \leq 0.065$), the cycle time of Model 1 is not long enough to absorb the increased costs of restoration and lost production, and therefore, Model 2 performs better where $0.776 \leq A \leq 0.800$. Low values of λ_1 and λ_2 (i.e., $0.0025 \leq \lambda_1 \leq 0.015$ and $0.0225 \leq \lambda_2 \leq 0.035$) enable the process to stay longer in the in-control state. This allows enough time to detect a shift before the occurrence of the other shift and reduces $LRCR$ of both models. The long in-control times in both models make $LRCR_{M2} \approx LRCR_{M1}$. As seen in Table 5, there is a noticeable increase in each model's $LRCR$ as λ_1 and λ_2 increase. For instance, $LRCR_{M1}$ of the first scenario is 14.86% and 52% less than $LRCR_{M1}$ of the original setup (i.e., $\lambda_1 = 0.01$ and $\lambda_2 = 0.03$) and the last scenario, respectively. Since the shift rate is one of the features of a machine, the decision maker can focus on how to reduce the shift rate. Redesigning or replacing machines to achieve a cost reduction could be a valuable option. For example, an automated painting chamber can be reinsulated with better insulation material to avoid spraying products with high viscous paint in a cold environment that reduces undesirable coating.

Table 5. Effect of shift parameters on the optimal solutions of the two models.

		Model 1				Model 2				
λ_1	λ_2	r	N	h	$LRCR_{M1}$	r	N	h	$LRCR_{M2}$	A
0.0025	0.0225	0	2	0.823	120.57*	0	2	0.840	120.76	0.800
0.005	0.025	0	2	0.832	128.20*	0	2	0.867	128.58	0.800

0.01	0.03	1	5	0.428	141.61	1	5	0.457	141.21*	0.800
0.015	0.035	1	5	0.435	152.78	1	5	0.481	151.71*	0.800
0.02	0.04	1	5	0.440	163.49	1	5	0.506	161.67*	0.800
0.025	0.045	1	5	0.448	173.42	1	4	0.325	171.16*	0.795
0.03	0.05	1	4	0.291	182.54	1	4	0.359	177.53*	0.785
0.035	0.055	0	12	4.085	194.71	1	4	0.363	185.97*	0.783
0.04	0.06	0	12	4.080	201.20	0	16	5.190	197.11*	0.782
0.045	0.065	0	12	4.073	207.69	0	16	4.688	206.27*	0.776
0.05	0.07	0	12	4.067	214.11*	0	16	4.275	215.40	0.767
0.055	0.075	0	12	4.059	220.43*	0	16	3.930	224.45	0.758
0.06	0.08	0	12	4.052	226.63*	0	16	3.637	233.39	0.750
0.065	0.085	0	12	4.044	232.70*	0	17	3.872	237.08	0.750
0.07	0.09	0	12	4.036	238.61*	0	17	3.620	245.35	0.742
0.075	0.095	0	12	4.028	244.37*	0	17	3.401	253.80	0.728
0.08	0.1	0	12	4.020	249.99*	0	17	3.206	261.93	0.724

Effect of C_{FA} on models' performances. As shown in Table 6, there is no significant difference between $LRCR_{M1}$ and $LRCR_{M2}$ at each level of C_{FA}, so either of the two models can be used. Naturally, the expected cost of false alarm increases as C_{FA} increases with the same sampling parameters. When C_{FA} >150, r increases to avoid frequent false alarms by accepting nonconforming units during inspection. Moreover, N increases to reduce type I error α and to achieve the desired PR_{eff}. Since with $r = 0$ and $N = 2$, α becomes high, the only way to reduce the number of false alarms is to reduce the number of samples taken by having a longer h. This justifies why h is higher for $C_{FA} \leq 150$ (Model 1) and $C_{FA} = 50$ (Model 2), and why it is lower for the other levels of C_{FA}. As seen in Table 6, there are two setups that can be used for inspection: for $C_{FA} < 200$ (Model 1), the setup with $(r, N, h) =$ (0, 2, 0.847) is appropriate, and for $C_{FA} \geq 200$, the setup with (1, 5, 0.428) is more economical. For Model 2, the setup with (0, 2, 0.838) is appropriate for $C_{FA} = 50$, whereas (1, 5, 0.457) is used for C_{FA} >50. Practitioners can choose between the two setups for a given value of C_{FA} without the need for solving the problem again (i.e., the two setups are usable for a wide range of C_{FA}). In addition, more solutions can be obtained from those setups by changing the decision variables slightly to achieve further reduction in $LRCR$ especially if the constraints are not violated significantly. This strategy allows more flexibility in selecting the most appropriate solution to cope with possible uncertainties and specific conditions. For instance, if a product is produced for a new customer, management may decide to reduce h (in Model 1) slightly to 0.800 as opposed to 0.847 ($C_{FA} <$ 200) to increase customer satisfaction by increasing the inspection frequency regardless of the increase in $LRCR_{M1}$.

Table 6. Effect of C_{FA} on the optimal solutions of the two models.

Model 1	Model 2

C_{FA}	r	N	h	$LRCR_{M1}$	r	N	h	$LRCR_{M2}$
50	0	2	0.841	131.54*	0	2	0.838	132.43
100	0	2	0.847	135.16*	1	5	0.457	136.07
150	0	2	0.847	138.78	1	5	0.457	138.64*
200	1	5	0.428	141.61	1	5	0.457	141.21*
250	1	5	0.428	144.19	1	5	0.457	143.77*
300	1	5	0.428	146.86	1	5	0.457	146.33*
350	1	5	0.428	149.42	1	5	0.457	148.90*

Effect of C_{LP} on models' performances. As seen in Table 7, there is no significant difference between $LRCR_{M1}$ and $LRCR_{M2}$ at each level of C_{LP}, and either of the two models can be used. Since the total cost increases with the increase in non-productive times such as sampling and false alarms, a high C_{LP} decreases N and increases h in order to increase AV. A low N means less time will be spent at each sampling, and a high h means a smaller number of samples will be taken, and hence, resulting in higher AV. On the contrary, a low C_{LP} permits to inspect more units but with a lower h. The higher values of N, as in the first scenario, reduce the number of false alarms by accepting nonconforming units during inspection ($r = 1$), and a low h reduces the cost of rejected units received by customers. For Model 1, practitioners can choose the setup with (0, 2, 0.847) for any $C_{LP} \geq 4$ and (1, 5, 0.428) for any $C_{LP} < 4$. For Model 2, the setup with (0, 2, 0.914) can be used for $C_{LP} \geq 7$, whereas (1, 5, 0.457) is appropriate for $C_{LP} \leq 6$. Hence, given the value of C_{LP}, the corresponding setup can be immediately identified for each model.

Table 7. Effect of C_{LP} on the optimal solutions of the two models.

	Model 1				Model 2			
C_{LP}	r	N	h	$LRCR_{M1}$	r	N	h	$LRCR_{M2}$
1	1	5	0.428	105.71*	1	5	0.457	106.13
2	1	5	0.428	123.67*	1	5	0.457	123.67
3	1	5	0.428	141.61	1	5	0.457	141.21*
4	0	2	0.847	159.22	1	5	0.457	158.74*
5	0	2	0.847	176.04*	1	5	0.457	176.27
6	0	2	0.847	192.84*	1	5	0.457	193.81
7	0	2	0.847	209.66*	0	2	0.914	210.97

Influence of ATS constraint L on models' performances. Table 8 illustrates the optimal solutions of the two models under different levels of L. A high L allows the process to operate for a long time without alerting a true alarm. This increases the total cost and cycle length of the two models. Because Model 2 allows only one shift to occur, the increase in its cycle length is much less compared to that of Model 1. For instance, when $L = 13.95$, the cycle length of Model 1 is 27.04% longer than that of Model 2. This makes Model 1 more economical than Model 2 for $L \geq 9.5$. For $L \leq 9$, $LRCR_{M2}$ on average is just 0.64% less than $LRCR_{M1}$, whereas $LRCR_{M1}$ is 1.92% less than $LRCR_{M2}$ for $L \geq 9.5$. It is worth pointing that

$LRCR_{M1}$ approaches a constant when L >13.50, and $LRCR_{M2}$ approaches a constant when $L \geq 9.50$. This means that relaxing the constraint on ATS makes Model 1 preferable than Model 2 under $AV \geq 0.8$ and $PR_{eff} \geq 0.9$. Clearly, further reductions in $LRCR_{M1}$ and $LRCR_{M2}$ can be achieved if L is increased from 3 to 13.95 while keeping other constraints unviolated. If more interest is in signaling an earlier true alarm, L can be further reduced down to 2 without affecting other constraints but increasing $LRCR_{M1}$ and $LRCR_{M2}$. Any increment for L >13.95 violates the constraint on PR_{eff}, whereas the constraint on AV is violated for $L < 2$.

Table 8. Influence of ATS on the optimal solutions of the two models.

	Model 1				Model 2				
L	r	N	h	$LRCR_{M1}$	r	N	h	$LRCR_{M2}$	A
0.5	0	2	0.138	245.16	0	2	0.141	243.88*	0.583
1	0	2	0.281	187.37	0	2	0.287	186.76*	0.725
1.5	0	2	0.422	163.37	0	2	0.437	162.80*	0.784
2	0	2	0.560	151.56	0	2	0.591	151.07*	0.800
2.5	0	2	0.700	145.42*	0	2	0.750	145.42*	0.800
3	1	5	0.428	141.61	1	5	0.457	141.20*	0.800
3.5	1	5	0.500	137.09	1	5	0.541	136.80*	0.800
4	1	4	0.367	133.71	1	4	0.402	132.82*	0.800
4.5	1	4	0.415	130.27	1	4	0.460	129.53*	0.800
5	1	4	0.462	127.68	1	4	0.519	127.15*	0.800
5.5	1	4	0.510	125.73	1	4	0.581	125.51*	0.800
6	1	4	0.558	124.29	1	3	0.345	122.94*	0.800
6.5	1	3	0.326	122.47	1	3	0.380	120.72*	0.800
7	1	3	0.352	120.53	1	3	0.417	118.86*	0.800
7.5	1	3	0.378	118.87	1	3	0.455	117.35*	0.800
8	1	3	0.404	117.45	1	3	0.494	116.15*	0.800
8.5	1	3	0.431	116.23	1	3	0.536	115.21*	0.800
9	1	3	0.457	115.20	1	3	0.578	114.52*	0.800
9.5	1	3	0.485	114.32*	1	3	0.584	114.45	0.800
10	1	3	0.511	113.58*	1	3	0.584	114.45	0.800
10.5	1	3	0.539	112.96*	1	3	0.584	114.45	0.800
11	1	3	0.566	112.46*	1	3	0.584	114.45	0.800
11.5	1	3	0.593	112.05*	1	3	0.584	114.45	0.800
12	1	3	0.621	111.74*	1	3	0.584	114.45	0.800
12.5	1	3	0.649	111.51*	1	3	0.584	114.45	0.800
13	1	3	0.677	111.35*	1	3	0.584	114.45	0.800
13.5	1	3	0.705	111.26*	1	3	0.584	114.45	0.800
13.95	1	3	0.730	111.24*	1	3	0.584	114.45	0.800

The marginal effect of h. Figure 6 shows how the change in h affects $LRCR_{M1}$ and the performance measures when keeping other parameters unchanged. In Figure 6.a, AV increases as h increases up to 0.856, and then decreases as h goes beyond 0.856. Since ATS is a function of h and $ARL_{s_{12}}$, ATS is an increasing linear function of h for given values of r and N (constant $ARL_{s_{12}}$) as seen in Figure 6.b. In Figure 6.c, decreasing h increases inspection frequency and reduces the number of nonconforming units produced between two inspections, and hence, PR_{eff} increases. Figure 6.d shows that $LRCR_{M1}$ significantly decreases to the minimum value 130.21 at $h = 0.856$ by violating the constraint on ATS, and then, it slowly increases. If more interest is in reducing $LRCR_{M1}$, h can be increased beyond the optimal $h^* = 0.428$ by violating some constraints. This may be satisfying if the violations are not significant. For instance, with h= 0.856, $LRCR_{M1}$ reduces to 130.21, but ATS increases to 5.

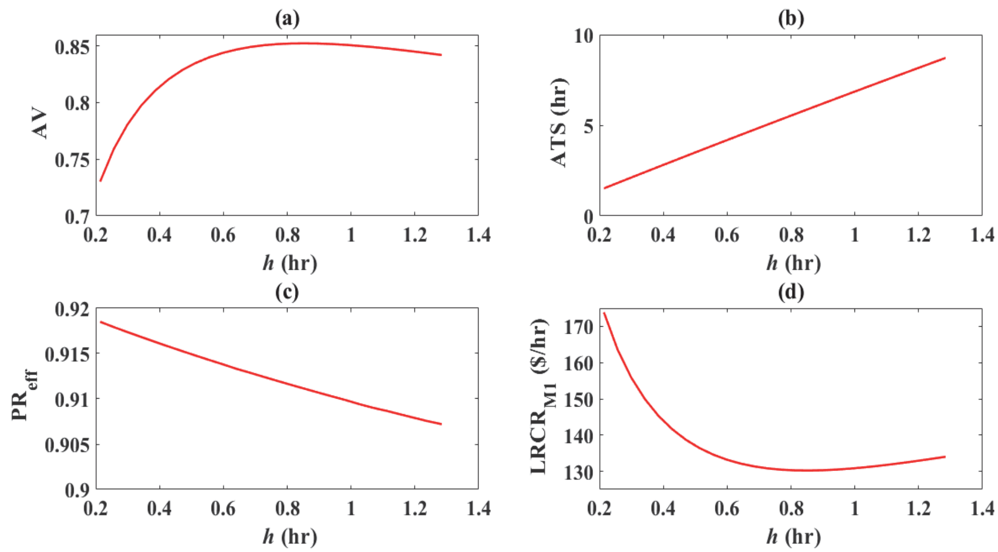

Figure 6. The marginal effect of h when r =1, N = 5.

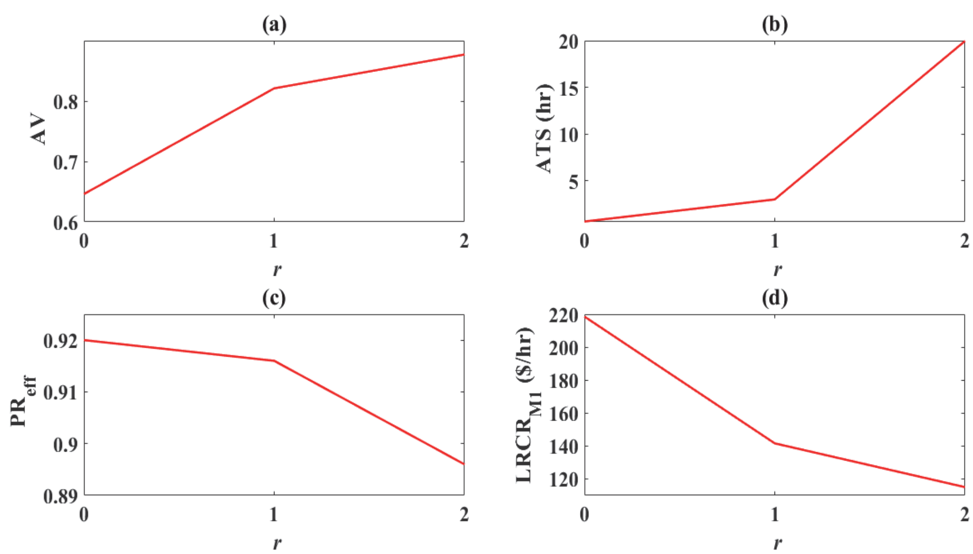

Figure 7. The marginal effect of r when h = 0.428, N = 5.

The marginal effect of r. Compared to the optimal setting $r^* = 1$, AV drops to 0.650 and ATS decreases to 0.63 at $r = 0$ as seen in Figures 7.a and 7.b, respectively. As r increases with respect to fixed N, the probability of missed detection (type II error) increases, and hence, ATS increases quite fast as shown in Figure 7.b. Moreover, PR_{eff} decreases as illustrated in Figure 7.c, as more nonconforming units are produced. Having $r = 0$, the corresponding number of false alarms is about 8 and 70 times the numbers of false alarms for $r = 1$ and $r = 2$, respectively. This drastically increases $LRCR_{M1}$ to 219 due to poor AV as depicted in Figure 7.d. Basically, r is not flexible to change compared to h, as changing r causes significant violations on the constraints. Therefore, attention should be paid when changing the value of r.

The marginal effect of N. In Figure 8.b, ATS has a noticeable increase when N decreases to 4 and 3, then it slowly decreases as N goes to 6 and 7. Since ATS increases with the increase of h and/or $ARL_{s_{12}}$, a low N increases type II error given fixed r, and hence, $ARL_{s_{12}}$ increases. In Figure 8.c, PR_{eff} increases with the increase in N. As N increases, type II error decreases, and a smaller number of nonconforming units are produced. The linear trends in Figures 8.a and 8.d are expected since as N increases, the times and costs of inspection and false alarms increase causing $LRCR_{M1}$ to increase and AV to decrease. Like h, N is flexible to change for a benefit to some extent. For instance, $LRCR_{M1}$ can be reduced to 130 if ATS is violated and increased to 4.7 when N is reduced to 4. In addition, N can be increased to 6 in order to reduce ATS to less than 2.5 hours resulting in a slight decrease in AV but an increase in $LRCR_{M1} \approx 150$.

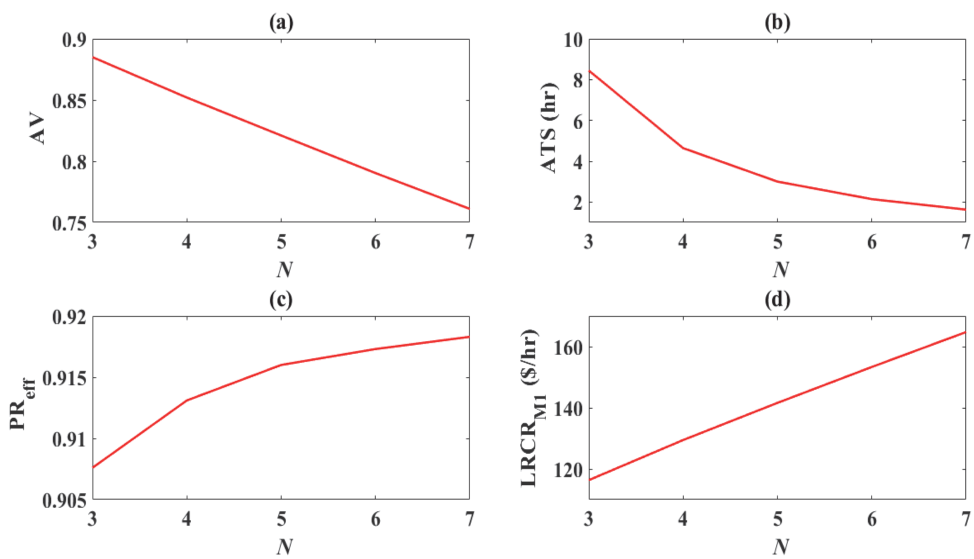

Figure 8. The marginal effect of N when $h = 0.428$, $r = 1$.

Practical guidelines for using the proposed model

As shown previously, allowing competing shifts to occur and propagate achieves some economic benefits for different settings of parameters. However, the number of scenarios for stochastic cases explained in Section 4 increases as the number of machines increases. In the current model, only two machines are considered with two scenarios ($T_1 > T_2$ & $T_1 > T_2$) for each case. If the number of machines increases to three, the total number of scenarios increases to 45. To make the model easier to handle, some assumptions can be made based on some prior understandings about the system. For instance, the model can be designed by allowing a certain number of shifts to occur, and such shifts cannot occur in the same sampling interval. Under this assumption, only scenarios of Case III need to be considered, and the three-machine system can be modelled with 12 scenarios instead of 45. Practitioners need to compromise between the economic benefits of considering propagating shifts and the design complexity. The proposed model can be used for systems with a larger number of machines by grouping machines into two aggregate stages. Within a stage, a combined effect (e.g., aggregate PON or shift rate) of machines can be considered instead of dealing with each machine alone. For instance, the shift of any machine (or all machines) in a stage may be assumed to have the same PON. This approach can also be applied to a machine where the degradation processes of different components cause quality deterioration (e.g., degradation of turbine discs and circulation mechanisms in the blasting machines). Apparently, combining stages reduces not only the number of stochastic scenarios but also the number of model parameters in a real-world application.

7. Conclusion and future work

Most of online sampling studies investigating multiple assignable causes are conducted on single-stage system. A few studies consider the multiplicity of assignable causes in multistage systems. However, those studies assume identical stages, $\bar{X}$ control chart, same shift level, economic model, no failures, or no quality related costs. This paper presents a sampling plan for attributes for a serial production system consisting of two unreliable machines where each machine is subject to sudden failure and shift in quality. A comprehensive economic-statistical model is developed to investigate the joint effect of different shifts by considering the stochastic competency and propagation of the shifts during manufacturing. The developed model generalizes all previous works and compromises between the quality and the quantity performances. The proposed sampling plan minimizes the long-run cost rate subject to constraints on system availability, effective production rate, and average time to signal. A thorough analysis is conducted on some input parameters, the constraint on average time to signal, and the marginal effects of decision variables. Specially, investigating the effects of process parameters, such as shift rates, helps management take long-term decisions (e.g., system overhaul and replacement). The analysis shows that when some decision variables are flexible to change, some adjustments can be made to emphasize specific needs. More importantly, compared to an alternative design that allows for only one assignable cause to occur in a single-stage system with multiple assignable causes, the proposed design shows better economic performance under different problem settings.

It is worth pointing out that since this work assumes that sampling is implemented at the end of a production line, the proposed sampling plan can handle a single-stage system (e.g., one machine) with multiple assignable causes and shift propagation by setting all the machines to be the same (i.e., identical machines with the same failure rates). In other words, such a single-stage system is a special case of our unreliable multistage system subject to competing and propagating random shifts, and it cannot be used when assignable causes are attributed to different machines with different failure rates.

There are some situations where the assumptions given in Section 3 do not hold. First, if the production rates and reliability of the two machines are significantly different and there are limited areas for storing WIP, the faster and the more reliable machine may have to be stopped to reduce WIP for lowering the related inventory costs. Then, issues with starving and blocking arise. As a result, the developed model in this work is unsuitable, and a new model must be developed to include decisions on the buffer size and inventory control. Second, if the two machines are dependent (i.e., a failure or a shift of one machine affects the other), a more complex model and different maintenance strategies are needed. Third, to avoid producing more nonconforming units, we assume the system will be preventively stopped during sampling. This is worthwhile if the sampling interval is long (the chance for the system to have a shift is high) and measuring the sampled units takes a while. If the production is allowed to continue during sampling, a delay time due to searching for a true alarm must be added to the average time to signal, and an additional cost due to potentially producing more nonconforming units must be considered. Beyond these, this work can be extended in other directions. In particular, a multistage system with more than two machines can be considered. Moreover, more than two states of product quality and multiple deterioration states of each machine can be considered. Clearly, the number of system states exponentially increases as the number of machines and/or the number of states of each machine get bigger. For such a complex situation, a simulation-based optimization approach may be utilized. In addition, some practical guidelines for using the model are illustrated. Finally, other system configurations, such as a series-parallel system and parallel-series system, can be studied to deal with cases involving multiple identical machines that perform the same actions during production.

Acknowledgments

The authors would like to thank the Department Editor, Associate Editor and two reviewers for their insightful comments and suggestions that greatly improved the quality of this paper.

Funding

This research was partly supported by the U.S. National Science Foundation (Grant #CMMI 1635379).

Notes on contributors

Dr. Sinan Obaidat received his B.S. and M.S. degrees in Industrial Engineering from Jordan University of Science & Technology and Ph.D. degree in Industrial Engineering from the University of Arkansas – Fayetteville, USA. He is currently an Assistant Professor of the Department of Industrial Engineering at Yarmouk University, Jordan. His research interest is in the area of decision modeling of maintenance, quality, and reliability with applications in production systems.

Dr. Haitao Liao is a Professor and John and Mar Lib White Endowed Systems Integration Chair in the Department of Industrial Engineering at the University of Arkansas – Fayetteville. He received a Ph.D. degree in Industrial and Systems Engineering from Rutgers University in 2004. He also earned M.S. degrees in Industrial Engineering and Statistics from Rutgers University, and a B.S. degree in Electrical Engineering from Beijing Institute of Technology. His research interests include: (1) reliability models, (2) maintenance and service logistics, (3) prognostics, (4) probabilistic risk assessment, and (5) analytics of sensor data. His research has been sponsored by the National Science Foundation, Department of Energy, Nuclear Regulatory Commission, Oak Ridge National Laboratory, and industry. The findings of his group have been published in IISE Transactions, European Journal of Operational Research, Naval Research Logistics, IEEE Transactions on Reliability, IEEE Transactions on Cybernetics, The Engineering Economist, Reliability Engineering & System Safety, etc. He received a National Science Foundation CAREER Award in 2010, IISE William A. J. Golomski Award in 2011, 2014 and 2018, SRE Stan Ofsthun Best Paper Award in 2015 and 2019, and 2017 Alan O. Plait Award for Tutorial Excellence. He is a Fellow of IISE, a member of INFORMS, and a lifetime member of SRE.

References

Ahmed, I., Sultana, I., Paul, S. K. and Azeem, A. (2014) Performance evaluation of control chart for multiple assignable causes using genetic algorithm. *International Journal of Advanced Manufacturing Technology*, **70**, 1889–1902.

Abolmohammadi, M., Seif, A., Behzadi, M. H. and Moghadam, M. B. (2019) Economic statistical design of adaptive $\overline{X}$ control charts based on quality loss functions. *Operation Research*, Forthcoming.

Azadeh, A., Sangari, M.S., Sangari, E. and Fatehi, S. (2015) A particle swarm algorithm for optimising inspection policies in serial multistage production processes with uncertain inspection costs. *International Journal of Computer Integrated Manufacturing,* **28** (7), 766-780.

Bai, D. S., and Yun, H. J. (1996) Optimal allocation of inspection effort in a serial multistage production system. *Computers & Industrial Engineering,* **30** (3), 387-396.

Bouslah, B., Gharbi, A. and Pellerin, R. (2013) Joint optimal lot sizing and production control policy in an unreliable and imperfect manufacturing system. *International Journal of Production Economics*, **144**(1), 143-156.

Cao,Y. and Subramaniam, V. (2013) Improving the performance of manufacturing systems with continuous sampling plans. *IIE Transactions,* **45**(6), 575-590.

Charongrattanasakul, P. and Pongpullponsak, A. (2011) Minimizing the cost of integrated systems approach to process control and maintenance model by EWMA control chart using genetic algorithm. *Expert Systems with Applications,* **38**(5), 5178-5186.

Colledani, M. and Tolio , T. (2012) Integrated quality, production logistics and maintenance analysis of multistage asynchronous manufacturing systems with degrading machines. *CIRP Annals-Manufacturing Technology,* **61**(1), 455-458.

Colledani, M. and Tolio , T. (2012) Integrated quality, production logistics and maintenance analysis of multistage asynchronous manufacturing systems with degrading machines. *CIRP Annals-Manufacturing Technology,* **61**(1), 455-458.

Kim, J., Jeong, M. K. and Elsayed, E.A. (2017) Monitoring multistage processes with autocorrelated observations. *International Journal of Production Research,* **55** (8), 2385–2396.

Kim, J. and Gershwin S. B. (2005) Integrated quality and quantity modeling of a production Line. *OR Spectrum,* **27**, 287-314.

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

Lam, Y. C., M., S., Zhang, S. and Wu, Z. (2005) Integrated control chart system - optimization of sample Sizes, sampling intervals and control limits. *International Journal of Production Research,* **43** (3), 563-582.

Liberopoulos, G., Kozanidis, G. and Tsarouhas, P. (2007) Performance evaluation of an automatic transfer line with WIP scrapping during long failures. *Manufacturing & Service Operations Management,* **9** (1), 62-83.

Linderman, K., McKone-Sweet, K.E. and Anderson, J.C. (2005) An integrated system approach to process control and maintenance. *European Journal of Operational Research,* **164**(2), 324-340.

Liu, L., Yu, M., Ma, Y. and Tu, Y. (2013) Economic and economic-statistical designs of an $\bar{X}$ control chart for two-unit series systems with condition-based maintenance. *European Journal of Operational Research,* **226**(3), 491-499.

Lorenzen, T.J. and Vance, L.C. (1986) The economic design of control charts: A unified approach. *Technometrics,* **28**(1), 3-10.

Malaki, M., Niaki, S. T. A. and Ershadi, M. J. (2011) A comparative study of four evolutionary algorithms for economic and economic-statistical designs of MEWMA control charts. *Journal of Optimization in Industrial Engineering,* **9**, 1-13.

MATLAB & Simulink. (2019) Mixed Integer Optimization.

https://www.mathworks.com/help/gads/mixed-integer-optimization.html#bs1cifg. Accessed October 22, 2019.

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

Mehrafrooz, Z. and Noorossan, R. (2011) An integrated model based on statistical process control and maintenance. *Computers & Industrial Engineering,* **61**(4),1245-1255.

Moghadam, M. B., Khadem, Y., Fani, S. and Pasha, M. A. (2018) Effects of non-normality on economic and economic statistical designs of $\bar{X}$ control charts with multiple assignable causes and Weibull in-control times. *Communications in Statistics-Simulation and Computation,* **47** (7), 2055-2069.

Mohammadi, M., Dantan, J. Y., Siadat, A. and Tavakkoli-Moghaddam, R. (2018) A bi-objective robust inspection planning model in a multi-stage serial production system M. *International Journal of Production Research,* **56** (4), 1432-1457.

Montgomery, D.C. (2009). *Introduction to Statistical Quality Control*. 6th ed. Wiley: Hoboken, N.J.

Naebulharama, R. Zhangb, L. (2014) Bernoulli serial lines with deteriorating product quality: performance evaluation and system-theoretic properties. *International Journal of Production Research,* **52**(5),1479-1494.

Pandey, D., Kulkarni, M.S. and Vrat, P. (2011) A methodology for joint optimization for maintenance planning, process quality, and production scheduling. *Computers and Industrial Engineering*, **61**(4), 1098-1106.

Pasha, M. A., Moghadam, M. B., Fani, S. and Khadem, Y. (2018) Effects of quality characteristic distributions on the integrated model of Taguchi's loss function and economic statistical design of $\bar{\mathrm{X}}$-control charts by modifying the Banerjee and Rahim economic model. *Communications in Statistics - Theory and Methods,* **47**(8), 1842–1855.

Rau, H. and Chu, Y.-H. (2005) Inspection allocation planning with two types of workstation: WVD and WAD. *International Journal of Advanced Manufacturing Technology,* **25** (9-10), 947-953.

Rezaei-Malek, M., Mohammadi, M., Jean-Yves, Siadat, A. and Tavakkoli-Moghaddam, R. (2019) A review on optimisation of part quality inspection planning in a multi-stage manufacturing system. *International Journal of Production Research,* **57** (15-16), 4880-7897.

Safaei, A. S., Kazemzadeh, R. B. and Gan, H. S. (2015) Robust economic-statistical design of $\bar{\mathrm{X}}$ control chart. *International Journal of Production Research* **53**(14), 4446–4458.

Salmasnia, A., Abdzadeh, B. and Namdar, M. (2017) A joint design of production run length, maintenance policy and control chart with multiple assignable causes. *Journal of Manufacturing Systems,* **42**, 44-56.

Shi, J., and Zhou, S. (2009) Quality control and improvement for multistage systems : A survey. *IIE*

Transactions, **41**, 744-753.

Sultana, I., Ahmed, I., Chowdhury, A. H. and Paul, S. K. (2014) Economic design of $\bar{X}$ control chart using genetic algorithm and simulated annealing algorithm. *International Journal of Productivity and Quality Management*, **14** (3), 352-372.

Xiang, L. and Tsung, F. (2008) Statistical monitoring of multi-stage processes based on engineering models. *IIE Transactions*, **40**(10), 957-970.

Xiang, Y. (2013) Joint optimization of $\bar{X}$ control chart and preventive maintenance policies: A Discrete-Time Markov Chain approach. *European Journal of Operational Research,* **229**(2), 382-390.

Yang,Y.M., Su, C.Y. and Pearn,W.L. (2010) Economic design of $\bar{X}$ control charts for continuous flow process with multiple assignable causes. *International Journal of Production Economics,* **128**(1),110-117.

Yu, F.J. and Hou, J. L. (2006) Optimization of design parameters for $\bar{X}$ control charts with multiple assignable causes. *Journal of Applied Statistics,* **33**(3), 279-290.

Yu, F.J., Tsou, C.S., Huang, K.I. and Wu, Z. (2010) An economic-statistical design of $\bar{X}$ control charts with multiple assignable causes. *Journal of Quality,* **17**(4), 327-338.

Zantek, P. F.,Wright, G. P. and Plante, R. D. (2002) Systems with correlated stages process and product improvement in manufacturing systems with correlated stages. *Management Science,* **48**(5), 591-606.

Zhong, J. and Ma, Y. (2017) An integrated model based on statistical process control and maintenance for two-stage dependent processes.*Communications in Statistics: Simulation and Computation,* **46**(1), 106-126.

Zhou, S., Huang, Q. and Shi, J. (2003) State space modeling of dimensional variation propagation in multistage machining process using differential motion vectors. *IEEE Transactions on Robotics and Automation,* **19**(2), 296-309.

Zhu, H., Zhang, C. and Deng, Y. (2016) Optimization design of attribute control charts for multi-station manufacturing system subjected to quality shifts. *International Journal of Production Research,* **54**(6), 1804-1821.

Appendix

Derivation of $E[V_{out}]$

In Case I, units are produced with $p_s = p_3$ .The expected number of samples taken until a true alarm is alerted is $ARL_{s_{12}}$ where $ARL_{s_{12}}$ is the average run length when the process is operating with S_{12}, and it is given in Montgomery (2009) as:

$$ARL_{s_{12}} = \frac{1}{1-\beta_{p_3}}.$$

The last sample which alerts the true signal has $r < d \leq N$. Hence, the expected number of rejected units found during sampling when the process is out-of-control given case I $E[V_{out}|\text{Case I}]$ is expressed as:

$$E[V_{out}|\text{Case I}] = \{a_{p_3} + (ARL_{s_{12}} - 1)b_{p_3}\},$$

where $ARL_{s_{12}} - 1$ samples do not alert a true alarm.

In Cases II & III, at least one sample is taken with $p_s = p_2$ if $T_1 > T_2$ or with $p_s = p_1$ if $T_2 > T_1$. Let Q_{p_2} and Q_{p_1} be the number of samples taken with $p_s = p_2$, and $p_s = p_1$, respectively. Then $E[Q_{p_2}|\text{Case II}_{T_1>T_2}]$ and $E[Q_{p_1}|\text{Case II}_{T_2>T_1}]$ are given as follows, respectively:

$$E[Q_{p_2}|\text{Case II}_{T_1>T_2}] = \frac{\sum_{q=1}^{\infty}\sum_{i=1}^{\infty} q(e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih})(e^{-\lambda_1(i+q-1)h}-e^{-\lambda_1(i+q)h})\beta_{p_2}^q}{\sum_{q=1}^{\infty}\sum_{i=1}^{\infty} (e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih})(e^{-\lambda_1(i+q-1)h}-e^{-\lambda_1(i+q)h})\beta_{p_2}^q} = \frac{e^{\lambda_1 h}}{(e^{\lambda_1 h}-\beta_{p_2})},$$

$$E[Q_{p_1}|\text{Case II}_{T_2>T_1}] = \frac{\sum_{q=1}^{\infty}\sum_{i=1}^{\infty} q(e^{-\lambda_1(i-1)h}-e^{-\lambda_1 ih})(e^{-\lambda_2(i+q-1)h}-e^{-\lambda_2(i+q)h})\beta_{p_1}^q}{\sum_{q=1}^{\infty}\sum_{i=1}^{\infty} (e^{-\lambda_1(i-1)h}-e^{-\lambda_1 ih})(e^{-\lambda_2(i+q-1)h}-e^{-\lambda_2(i+q)h})\beta_{p_1}^q} = \frac{e^{\lambda_2 h}}{(e^{\lambda_2 h}-\beta_{p_1})},$$

where q denotes to the number of samples taken between the occurrence times of S_1 and S_2. In Case III, $S_2(S_1)$ is always detected before the occurrence of $S_1(S_2)$, and hence, $E[Q_{p_2}|\text{Case III}_{T_1>T_2}]$ and $E[Q_{p_1}|\text{Case III}_{T_2>T_1}]$ are given as follows:

$$E[Q_{p_2}|\text{Case III}_{T_1>T_2}] = \frac{\sum_{w=1}^{\infty}\sum_{i=1}^{\infty} w(e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih})e^{-\lambda_1(i+w-1)h}\beta_{p_2}^{w-1}(1-\beta_{p_2})}{\sum_{w=1}^{\infty}\sum_{i=1}^{\infty} (e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih})e^{-\lambda_1(i+w-1)h}\beta_{p_2}^{w-1}(1-\beta_{p_2})} = \frac{e^{\lambda_1 h}}{(e^{\lambda_1 h}-\beta_{p_2})},$$

$$E[Q_{p_1}|\text{Case III}_{T_2>T_1}] = \frac{\sum_{w=1}^{\infty}\sum_{i=1}^{\infty} w(e^{-\lambda_1(i-1)h}-e^{-\lambda_1 ih})e^{-\lambda_2(i+w-1)h}\beta_{p_1}^{w-1}(1-\beta_{p_1})}{\sum_{w=1}^{\infty}\sum_{i=1}^{\infty} (e^{-\lambda_1(i-1)h}-e^{-\lambda_1 ih})e^{-\lambda_2(i+w-1)h}\beta_{p_1}^{w-1}(1-\beta_{p_1})} = \frac{e^{\lambda_2 h}}{(e^{\lambda_2 h}-\beta_{p_1})},$$

where w represents the number of samples that process undergoes with S_2 until a successful detection. The term $e^{-\lambda_1(i+w-1)h}$ indicates that S_2 is detected at the sampling time $(i+w-1)h$, at which, S_1 still has not occurred yet.Consequently, the expected number of rejected units during the inspection when the process is in the out-of-control period $E[V_{out}]$ can be obtained as:

$$E[V_{out}] = E[V_{out}|\text{Case I}]\{B_1 + B_2\} + E[Q_{p_2}|\text{Case II}_{T_1>T_2}]b_{p_2}B_3 + E[Q_{p_1}|\text{Case II}_{T_2>T_1}]b_{p_1}B_4$$
$$+ \{(ARL_{s_{12}} - 1)b_{p_3} + a_{p_3}\}\{B_3 + B_4\} + (E[Q_{p_2}|\text{Case III}_{T_1>T_2}] - 1)b_{p_2}B_5 +$$
$$(E[Q_{p_1}|\text{Case III}_{T_2>T_1}] - 1)b_{p_1}B_6 + a_{p_2}B_5 + a_{p_1}B_6,$$

where $\{B_1 + B_2\}$ is the total probability of Case I, B_3 is the probability of Case II given $T_1 > T_2$, B_4 is the probability of Case II given $T_2 > T_1$, B_5 is the probability of Case III given $T_1 > T_2$, and B_6 is the probability of Case III given $T_2 > T_1$. In the above equation, $E[Q_{p_2}|\text{Case II}_{T_1>T_2}]$ $(E[Q_{p_1}|\text{Case II}_{T_2>T_1}])$is the expected number of samples that don't alert a true alarm in Case II when a process operates with $S_2(S_1)$, $\left(ARL_{s_{12}}-1\right)$ is the average number of samples that don't alert a true alarm when the process operates with S_{12} in Case II, and a_{p_3}represents the expected number of rejected units in the last sample that alert a true alarm given Case II. In Case III, $E\left[Q_{p_2}\middle|\text{Case III}_{T_1>T_2}\right]-1\left(E\left[Q_{p_1}\middle|\text{Case III}_{T_2>T_1}\right]-1\right)$ is the expected number of samples that don't alert a true alarm when the process operates with $S_2(S_1)$, and $a_{p_2}\left(a_{p_1}\right)$ is the average number of rejected units found in the last sample that detects $S_2(S_1)$. B_1 to B_6 are given in Subsection 4.5, whereas $a_{p_s}, b_{p_s}, p_s \in \{p_0, p_1, p_2, p_3\}$ are given in Subsection 4.8.

Derivations of $E[T_{s_1}]$, $E[T_{s_2}]$, $E[T_{s_{12}}]$, u_1 to u_6, and C_1 to C_8

Case I. Given that S_2 and S_1 occur in the same sampling interval as shown in Figure 2, the conditional expectations of τ_{S_1} and τ_{S_2} are obtained as follows.

If $T_1 > T_2$, we have:

$$u_1 = E\left[\tau_{S_1}\middle|(i-1)h \le T_2 \le T_1 < ih\right] = \frac{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_1}(t_1-(i-1)h)\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1}dt_2dt_1}{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_1}\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1}dt_2dt_1}$$

$$= \frac{\lambda_2^{\,2}e^{\lambda_2 h}(e^{\lambda_1 h}-1)-\lambda_1^{\,3}h(e^{\lambda_2 h}-1)-\lambda_1\lambda_2 e^{\lambda_2 h}(2-2e^{\lambda_1 h}+\lambda_2 h)+\lambda_1^{\,2}(1+\lambda_2 h-e^{\lambda_2 h}(1+2\lambda_2 h))}{\lambda_1(\lambda_1+\lambda_2)(\lambda_1-e^{\lambda_2 h}(\lambda_1+\lambda_2-\lambda_2 e^{\lambda_1 h}))},$$

$$u_2 = E\left[\tau_{S_2}\middle|(i-1)h \le T_2 \le T_1 < ih\right] = \frac{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_1}(t_2-(i-1)h)\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1}dt_2dt_1}{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_1}\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1}dt_2dt_1}$$

$$= \frac{e^{\lambda_2 h}(\lambda_2^{\,2}e^{\lambda_1 h}-(\lambda_1+\lambda_2)^2)+\lambda_1(\lambda_1+2\lambda_2+\lambda_2(\lambda_1+\lambda_2)h)}{\lambda_2(\lambda_1+\lambda_2)(\lambda_1-e^{\lambda_2 h}(\lambda_1+\lambda_2-\lambda_2 e^{\lambda_1 h}))}.$$

Since S_2 occurs before S_1 in the same sampling interval, S_2 propagates to S_{12} at the time of S_1 occurrence and prior to the next sampling time. As a result, we have:

$$E\left[T_{s_2}\middle|\text{Case I}_{T_1>T_2}\right] = u_1 - u_2,\ E\left[T_{s_1}\middle|\text{Case I}_{T_1>T_2}\right] = 0,\ E\left[T_{s_{12}}\middle|\text{Case I}_{T_1>T_2}\right] = hARL_{s_{12}} - u_1,$$

where $ARL_{s_{12}}$ is the average run length when the system operates with S_{12}, i.e., with $p_s = p_3$. The average length in the out-of-control state is defined as the average number of samples taken since the occurrence of a shift until a true alarm is alerted.

The corresponding probability of Case I, $T_1 > T_2$ is:

$$C_1 = \sum_{i=1}^{\infty} P\left((i-1)h \le T_2 \le T_1 < ih\right) = \frac{\lambda_1\left(1-e^{\lambda_2 h}\right)+\lambda_2(e^{(\lambda_1+\lambda_2)h}-e^{\lambda_2 h})}{(\lambda_1+\lambda_2)(e^{(\lambda_1+\lambda_2)h}-1)},$$

where

$$P\big((i-1)h \le T_2 \le T_1 < ih\big) = \int_{(i-1)h}^{ih} \int_{(i-1)h}^{t_1} \lambda_2 e^{-\lambda_2 t_2}\, \lambda_1 e^{-\lambda_1 t_1} dt_2\, dt_1$$

$$= e^{-\lambda_2(i-1)h}\big(e^{-\lambda_1(i-1)h} - e^{-\lambda_1 ih}\big) + \frac{\lambda_1}{\lambda_1+\lambda_2}\big(e^{-(\lambda_1+\lambda_2)ih} - e^{-(\lambda_1+\lambda_2)(i-1)h}\big).$$

If $T_2 > T_1$, we have:

$$u_3 = E\big[\tau_{S_1}\big|(i-1)h \le T_1 \le T_2 < ih\big] = \frac{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_2}(t_1-(i-1)h)\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1} dt_1 dt_2}{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_2}\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1} dt_1 dt_2}$$

$$= \frac{e^{\lambda_1 h}({\lambda_1}^2 e^{\lambda_2 h} - (\lambda_1+\lambda_2)^2) + \lambda_2(\lambda_2 + 2\lambda_1 + \lambda_1(\lambda_1+\lambda_2)h)}{\lambda_1(\lambda_1+\lambda_2)(\lambda_2 - e^{\lambda_1 h}(\lambda_1+\lambda_2-\lambda_1 e^{\lambda_2 h}))},$$

$$u_4 = E\big[\tau_{S_2}\big|(i-1)h \le T_1 \le T_2 < ih\big] = \frac{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_2}(t_2-(i-1)h)\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1} dt_1 dt_2}{\int_{(i-1)h}^{ih}\int_{(i-1)h}^{t_2}\lambda_2 e^{-\lambda_2 t_2}\lambda_1 e^{-\lambda_1 t_1} dt_1 dt_2} =$$

$$\frac{{\lambda_1}^2 e^{\lambda_1 h}(e^{\lambda_2 h}-1) - {\lambda_2}^3 h(e^{\lambda_1 h}-1) - \lambda_1\lambda_2 e^{\lambda_1 h}(2 - 2e^{\lambda_2 h} + \lambda_1 h) + {\lambda_2}^2(1+\lambda_1 h - e^{\lambda_1 h}(1+2\lambda_1 h))}{\lambda_2(\lambda_1+\lambda_2)(\lambda_2 - e^{\lambda_1 h}(\lambda_1+\lambda_2-\lambda_1 e^{\lambda_2 h}))}.$$

Since S_1 occurs before S_2 in the same sampling interval, S_1 propagates to S_{12} at the time of S_2 occurrence and prior to the next sampling time. Therefore:

$$E\big[T_{S_2}\big|\text{Case I}_{T_2>T_1}\big] = 0,\ E\big[T_{S_1}\big|\text{Case I}_{T_2>T_1}\big] = u_4 - u_3,\ E\big[T_{S_{12}}\big|\text{Case I}_{T_2>T_1}\big] = hARL_{S_{12}} - u_4.$$

The corresponding probability of Case I, $T_2 > T_1$ is:

$$C_2 = \sum_{i=1}^{\infty} P\big((i-1)h \le T_1 \le T_2 < ih\big) = \frac{\lambda_2\big(1-e^{\lambda_1 h}\big) + \lambda_1(e^{(\lambda_1+\lambda_2)h} - e^{\lambda_1 h})}{(\lambda_1+\lambda_2)(e^{(\lambda_1+\lambda_2)h}-1)},$$

where

$$P\big((i-1)h \le T_1 \le T_2 < ih\big) = \int_{(i-1)h}^{ih} \int_{(i-1)h}^{t_2} \lambda_1 e^{-\lambda_1 t_1}\, \lambda_2 e^{-\lambda_2 t_2} dt_1\, dt_2$$

$$= e^{-\lambda_1(i-1)h}\big(e^{-\lambda_2(i-1)h} - e^{-\lambda_2 ih}\big) + \frac{\lambda_2}{\lambda_1+\lambda_2}\big(e^{-(\lambda_1+\lambda_2)ih} - e^{-(\lambda_1+\lambda_2)(i-1)h}\big).$$

Cases II & III. In Cases II and III, S_2 and S_1 occur in different sampling intervals as shown in Figures 3 and 4 where $0 \le \tau_{S_1} \le h$, and $0 \le \tau_{S_2} \le h$. Therefore, the conditional expectations of τ_{S_1} and τ_{S_2} are given as follows, respectively:

$$u_5 = E\big[\tau_{S_1}\big|(i-1)h \le T_1 < ih\big] = \frac{\int_{(i-1)h}^{ih}(t_1-(i-1)h)\lambda_1 e^{-\lambda_1 t_1}\, dt_1}{\int_{(i-1)h}^{ih}\lambda_1 e^{-\lambda_1 t_1}\, dt_1} = \frac{1-(1+\lambda_1 h)e^{-\lambda_1 h}}{\lambda_1(1-e^{-\lambda_1 h})},$$

$$u_6 = E\left[\tau_{S_2}\middle|(i-1)h \le T_2 < ih\right] = \frac{\int_{(i-1)h}^{ih}(t_2-(i-1)h)\lambda_2 e^{-\lambda_2 t_2}\,dt_2}{\int_{(i-1)h}^{ih}\lambda_2 e^{-\lambda_2 t_2}\,dt_2} = \frac{1-(1+\lambda_2 h)e^{-\lambda_2 h}}{\lambda_2(1-e^{-\lambda_2 h})}.$$

Cases II. $E[T_{s_1}]$ and $E[T_{s_2}]$ depend on how many samples $q, q = \{1,\cdots,\infty\}$ are between T_1 and T_2. For instance, if S_1 occurs three samples after the occurrence of S_2, then $E[T_{s_2}] = 3h - u_6 + u_5$ given that S_2 is not detected until the occurrence of S_1.

If $T_1 > T_2$, we have:

$$C_3 = E\left[T_{s_2}, \text{Case II}_{T_1>T_2}\right]$$
$$= \sum_{q=1}^{\infty}\sum_{i=1}^{\infty}(qh-u_6+u_5)\left(e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih}\right)\left(e^{-\lambda_1(i+q-1)h}-e^{-\lambda_1(i+q)h}\right)\beta_{p_2}^{q}$$
$$= \frac{\beta_{p_2}(e^{\lambda_1 h}-1)(e^{\lambda_2 h}-1)(e^{\lambda_1 h}(h+u_5-u_6)+\beta_{p_2}(u_6-u_5))}{(e^{(\lambda_1+\lambda_2)h}-1)\left(e^{\lambda_1 h}-\beta_{p_2}\right)^2},$$

$$E\left[T_{s_{12}}\middle|\text{Case II}_{T_1>T_2}\right] = hARL_{s_{12}} - u_5.$$

In C_3, S_2 occurs in the sampling interval $[(i-1)h, ih]$ and S_1 occurs in the sampling interval $[(i+q-1)h, (i+q)h]$ afterwards. For instance, if S_2 occurs in $[0, h]$, then S_1 could occur one sample afterwards, i.e., $[h, 2h]$, or two samples afterwards, i.e., $[2h, 3h]$, and so on. For any q, the sampling plan always fails to detect S_2 until the occurrence of S_1 resulting in $\beta_{p_2}^{q}$ type II error.

If $T_2 > T_1$, we have:

$$C_4 = E\left[T_{s_1}, \text{Case II}_{T_2>T_1}\right]$$
$$= \sum_{q=1}^{\infty}\sum_{i=1}^{\infty}(qh-u_5+u_6)\left(e^{-\lambda_1(i-1)h}-e^{-\lambda_1 ih}\right)\left(e^{-\lambda_2(i+q-1)h}-e^{-\lambda_2(i+q)h}\right)\beta_{p_1}^{q}$$
$$= \frac{\beta_{p_1}(e^{\lambda_2 h}-1)(e^{\lambda_1 h}-1)(e^{\lambda_2 h}(h+u_6-u_5)+\beta_{p_1}(u_5-u_6))}{(e^{(\lambda_1+\lambda_2)h}-1)\left(e^{\lambda_2 h}-\beta_{p_1}\right)^2},$$

$$E\left[T_{s_{12}}\middle|\text{Case II}_{T_2>T_1}\right] = hARL_{s_{12}} - u_6.$$

Cases III. If $T_1 > T_2$, then sampling plan is always able to detect S_2 before the occurrence of S_1 as shown in Figure 4. Therefore, the system is only operating with S_2. For instance, $E[T_{s_2}] = h - u_6$, if S_2 is immediately detected at the next sampling time and before the occurrence of S_1. $E[T_{s_2}] = 2h - u_6$, if S_2 is detected two sampling times since its occurrence and before the occurrence of S_1. Sampling fails to detect S_2 at the first sampling time, but it can detect it at the second sampling time. The following formula generalizes this situation:

$$C_5 = E\left[T_{s_2}, \text{Case III}_{T_1>T_2}\right] = \sum_{w=1}^{\infty}\sum_{i=1}^{\infty}(wh-u_6)\left(e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih}\right)e^{-\lambda_1(i+w-1)h}\beta_{p_2}^{w-1}\left(1-\beta_{p_2}\right)$$

$$= \frac{\left(1-\beta_{p_2}\right)e^{\lambda_1 h}\left(e^{\lambda_2 h}-1\right)\left(e^{\lambda_1 h}(h-u_6)+\beta_{p_2}u_6\right)}{\left(e^{(\lambda_1+\lambda_2)h}-1\right)\left(e^{\lambda_1 h}-\beta_{p_2}\right)^2},$$

where w represents the number of samples that process undergoes with S_2 until a success detection. The term $e^{-\lambda_1(i+w-1)h}$ indicates that S_2 is detected at the sampling time $(i+w-1)h$, at which, S_1 still has not occurred yet. For example, if S_2 occurs in the time interval $[h, 2h\,]$, then $E\left[T_{s_2}\,\right]=h-u_6$ if S_2 is detected at time $2h$, and hence, $i=2, w=1$, and

$$(wh-u_6)\left(e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih}\right)e^{-\lambda_1(i+w-1)h}\beta_{p_2}^{w-1}\left(1-\beta_{p_2}\right)$$
$$=(h-u_6)\left(e^{-\lambda_2 h}-e^{-\lambda_2 2h}\right)e^{-\lambda_1 2h}\left(1-\beta_{p_2}\right).$$

$E\left[T_{s_2}\,\right]=2h-u_6$ if S_2 is detected at time $3h$, and hence, $i=2, w=2$, and

$$(wh-u_6)\left(e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih}\right)e^{-\lambda_1(i+w-1)h}\beta_{p_2}^{w-1}\left(1-\beta_{p_2}\right)$$
$$=(2h-u_6)\left(e^{-\lambda_2 h}-e^{-\lambda_2 2h}\right)e^{-\lambda_1 3h}\beta_{p_2}\left(1-\beta_{p_2}\right).$$

If $T_2 > T_1$, then sampling plan is always able to detect S_1 before the occurrence of S_2. Therefore, the system is only operating with S_1. The same derivation approach like in $T_1 > T_2$ is followed, and hence:

$$C_6 = E\left[T_{s_1}, \text{Case III}_{T_2>T_1}\,\right] = \sum_{w=1}^{\infty}\sum_{i=1}^{\infty}(wh-u_5)\left(e^{-\lambda_1(i-1)h}-e^{-\lambda_1 ih}\right)e^{-\lambda_2(i+w-1)h}\beta_{p_1}^{w-1}\left(1-\beta_{p_1}\right)$$
$$= \frac{(1-\beta_{p_1})e^{\lambda_2 h}(e^{\lambda_1 h}-1)(e^{\lambda_2 h}(h-u_5)+\beta_{p_1}u_5)}{(e^{(\lambda_1+\lambda_2)h}-1)(e^{\lambda_2 h}-\beta_{p_1})^2}.$$

Note that there is no chance for propagating shift to occur in Case III, and therefore:

$$E\left[T_{s_{12}}, \text{Case III}_{T_1>T_2}\,\right] = E\left[T_{s_{12}}, \text{Case III}_{T_2>T_1}\,\right] = 0.$$

Considering all the above, $E[T_{s_1}]$, $E[T_{s_2}]$, and $E[T_{s_{12}}]$, are given as follows, respectively:

$$E[T_{s_1}] = \{u_4-u_3\}C_2 + C_4 + C_6,$$
$$E[T_{s_2}] = \{u_1-u_2\}C_1 + C_3 + C_5,$$
$$E[T_{s_{12}}] = \left\{hARL_{s_{12}}-u_1\right\}C_1 + \left\{hARL_{s_{12}}-u_4\right\}C_2 + \left\{hARL_{s_{12}}-u_5\right\}C_7 + \left\{hARL_{s_{12}}-u_6\right\}C_8,$$

where C_7 is the probability that the time needed is $hARL_{s_{12}}-u_5$ to alert a true alarm since the occurrence of a shift given Case II, $T_1 > T_2$, whereas C_8 is the probability that the time needed is $hARL_{s_{12}}-u_6$ to alert a true alarm since the occurrence of a shift given Case II, $T_2 > T_1$. C_7 and C_8 are given by:

$$C_7 = \sum_{q=1}^{\infty}\sum_{i=1}^{\infty}\left(e^{-\lambda_2(i-1)h}-e^{-\lambda_2 ih}\right)\left(e^{-\lambda_1(i+q-1)h}-e^{-\lambda_1(i+q)h}\right)\beta_{p_2}^{q}$$
$$= \frac{\beta_{p_2}e^{-(4\lambda_1+\lambda_2)h}\left(e^{\lambda_1 h}-1\right)\left(e^{\lambda_2 h}-1\right)\left(\beta_{p_2}e^{(4\lambda_1+2\lambda_2)h}-e^{(4\lambda_1+\lambda_2)h}\right)}{\left(e^{(\lambda_1+\lambda_2)h}-1\right)\left(e^{\lambda_1 h}-\beta_{p_2}\right)\left(\beta_{p_2}e^{\lambda_2 h}-1\right)},$$

$$C_8 = \sum_{q=1}^{\infty}\sum_{i=1}^{\infty}\left(e^{-\lambda_1(i-1)h} - e^{-\lambda_1 ih}\right)\left(e^{-\lambda_2(i+q-1)h} - e^{-\lambda_2(i+q)h}\right)\beta_1^q$$

$$= \frac{\beta_{p_1}e^{-(\lambda_1+4\lambda_2)h}(e^{\lambda_1 h}-1)(e^{\lambda_2 h}-1)(\beta_{p_1}e^{(2\lambda_1+4\lambda_2)h}-e^{(\lambda_1+4\lambda_2)h})}{(e^{(\lambda_1+\lambda_2)h}-1)(e^{\lambda_2 h}-\beta_{p_1})(\beta_{p_1}e^{\lambda_1 h}-1)}.$$