IE 5345 Project Final Report

Student: Ahsanul Abedin

ID: A05376947

Instructor: Dr. Clara Novoa

Multi-Objective Optimization of Redundancy, Maintenance, and Spare Inventory under Uncertainty.

Abstract

This project focuses on optimizing redundancy, maintenance, spare parts, and repair servers to reduce costs while enhancing system reliability and availability under uncertainty. A comprehensive system availability model, integrating ten performance factors, is developed as the foundation for the joint redundancy-inventory-repair allocation model with two stage stochastic programming. The model is applied to the automatic test equipment industry, where superior system reliability and operational availability are critical for global market competitiveness. Numerical analysis provides valuable managerial insights.

Keywords: TSSP; Multi Objective Optimization; Reliability; System Availability; installed base; decentralized repair; redundancy-maintenance-inventory model; superimposed renewal process.

Introduction:

1. Integrated Product Service Network

The system configuration is depicted in Figure 1. The studied system consists of multiple k_i-out-of-n_i active redundant subsystems (for i =1, 2, …, N) arranged in series. Within each subsystem, components are identical, but they vary across subsystems, resulting in N distinct component types. For the i^{th} subsystem, k_i represents the minimum number of operational units required, where $k_i \leq n_i$. As components are detachable and repairable, they are also termed line replaceable units (LRU). In this study, the terms component, and part are used interchangeably to denote a repairable LRU.

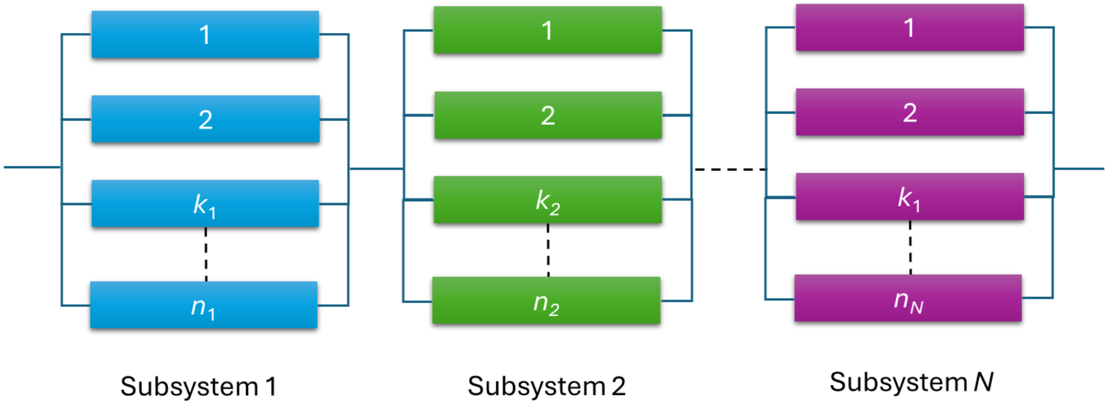
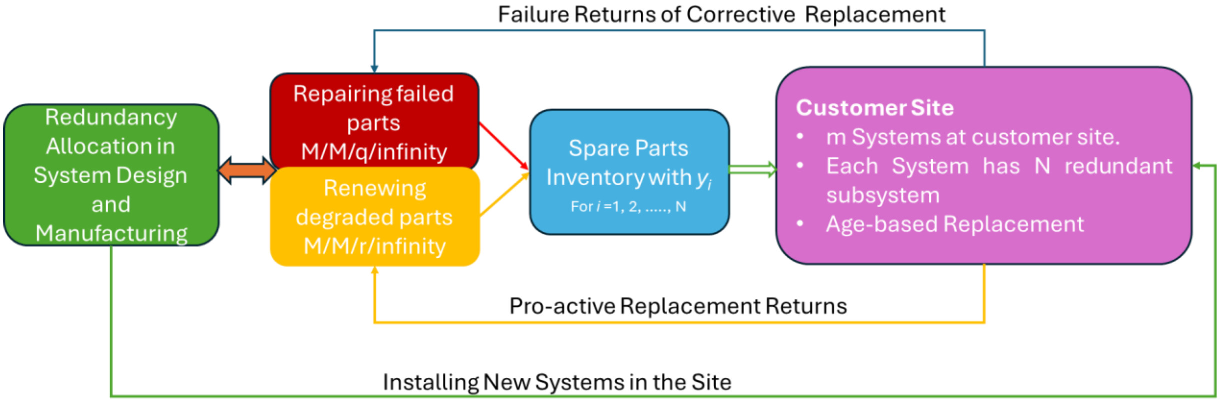


Figure 1: A System Comprised of N Redundant Subsystems in Series

As illustrated in Figure 2, the original equipment manufacturer (OEM) operates an integrated product-service program, overseeing design, production, and after-sales service/repair. At the customer site, m systems are installed. Due to the sporadic demand for spare parts, a spares inventory is strategically located near these systems to support replacements (Hekimoğlu et al.,


Failure Returns of Corrective Replacement
Redundancy Allocation in System Design and Manufacturing
Repairing failed parts M/M/q/infinity
Renewing degraded parts M/M/r/infinity
Spare Parts Inventory with y_i
For i =1, 2, ..., N
Customer Site
m Systems at customer site.
Each System has N redundant subsystem
Age-based Replacement
Pro-active Replacement Returns
Installing New Systems in the Site

2018). This model is common among OEMs such as semiconductor equipment suppliers, aircraft engine manufacturers, high-speed rail producers, and wind turbine makers, among others.

Figure 2: Product Service Integration with Decentralized Repair Services

In industry, age-based replacement is commonly adopted due to its technical reliability and scheduling versatility (El-Ferik, 2008; Huynh et al., 2012). For part type i (i =1, 2,…, N), inspections occur at a predetermined interval τ_i. If an item remains functional beyond τ_i, it is proactively replaced with a spare. If it fails before τ_i, an immediate corrective replacement is carried out. Consequently, the fleet generates two types of spare part demands: one for proactive replacements and another for failure-driven replacements. After renewal or repair, the part is returned to the inventory for future reuse.

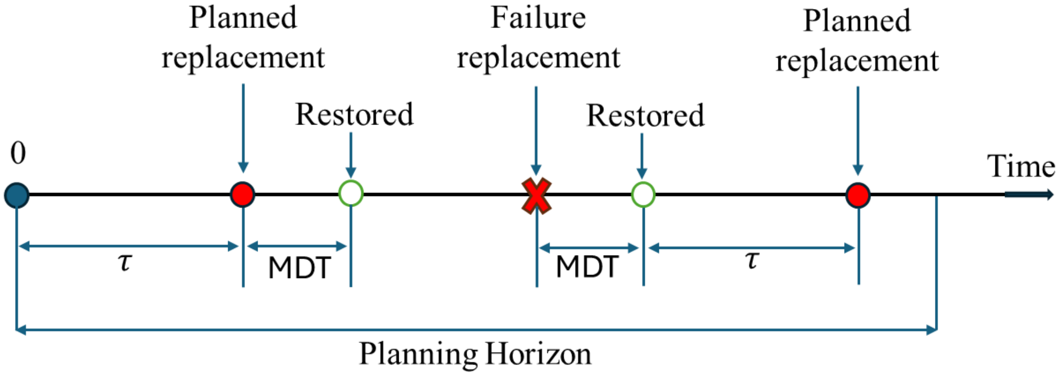


Figure 3: Working Principle of Age-based Maintenance (MDT= Mean Downtime)

2. Superimposed Renewal Process for Parts Replacement

Consider a single-item system with only one component. The reliability of this system under age-based maintenance is typically measured by the mean-time-between-replacements (MTBR). As the system consists of a single component, its failure results in system downtime. Let R(t) represent the component reliability and F(t) the cumulative distribution function. The MTBR can be calculated using a well-established formula in Reliability Engineering.

$$MTBR = \int_0^{\tau} R(t)dt = \tau - \int_0^{\tau} F(t)dt \, . \tag{1}$$

In age-based maintenance, the spare parts demand process can be treated as the superposition of two renewal processes: a proactive replacement stream and a failure (i.e., corrective) replacement stream (Jin et al. 2015). For a single-item system with one LRU, let $\lambda_p(\tau)$ and $\lambda_q(\tau)$ be the spare parts demand rate in proactive replacement and failure replacement, respectively. Based on Eq. (1), we have

$$\lambda_r(\tau) = \frac{R(\tau)}{\int_0^\tau R(t)dt}. \tag{2}$$

$$\lambda_q(\tau) = \frac{F(\tau)}{\int_0^\tau R(t)dt}. \tag{3}$$

Given a fleet with m single-item systems, each system independently generates proactive replacement and failure replacement streams, respectively. Hence the aggregate spare part demand rate of the fleet, denoted as $\lambda_m(\tau)$, can be estimated as

$$\lambda_m(\tau) = m \times \left(\lambda_p(\tau) + \lambda_q(\tau)\right) = \frac{m}{\int_0^\tau R(t)dt}. \tag{4}$$

The process formed by the union of fleet failures is called a *superimposed renewal process* (SRP). Cox and Smith (1954) have proved that, as m approaches the infinity and the operating time is large enough, SRP becomes a homogenous Poisson process regardless of the lifetime distribution of each system. In our study, since a failed part is swapped with a spare part, the replacement is equivalent to a perfect repair.

3. Parts Repair Queueing Model

SRP can be approximated as a homogeneous Poisson process if the fleet size $m \geq 10$. Thus Erlang-C queueing model can be used to characterize the repair shop performance. Let q be the number of repair servers, and $\lambda_{F,q}$ be the arrival rate of failed parts to the shop. If a fleet consists of m single-item systems, we have $\lambda_{F,q} = m\lambda_q(\tau)$ where $\lambda_q(\tau)$ is given in Eq. (3). The transition diagram of the $M/M/q/\infty$ queue is given in Figure 4.

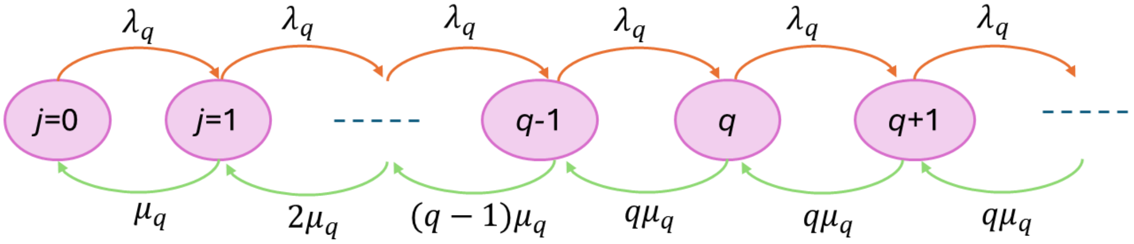


Figure 4: The $M/M/q/\infty$ Queueing Model for Parts Repair Process

The state of the queue represents the number of failed parts in the repair shop where μ_q is the repair rate of a server. Let $B(q)$ be the probability that an incoming part needs to wait in the queue. Then we have (Winston 2004):

$$B(q)=\frac{\dfrac{(\lambda_{F,q}/\mu_q)^q}{q!(1-\lambda_{F,q}/(q\mu_q))}}{\sum_{j=0}^{q-1}\dfrac{(\lambda_{F,q}/\mu_q)^j}{j!}+\dfrac{(\lambda_{F,q}/\mu_q)^q}{q!(1-\lambda_{F,q}/(q\mu_q))}}=\frac{\dfrac{(q\rho_q)^q}{q!(1-\rho_q)}}{\sum_{j=0}^{q-1}\dfrac{(q\rho_q)^j}{j!}+\dfrac{(q\rho_q)^q}{q!(1-\rho_q)}}, \qquad (5)$$

where $\rho_q=\lambda_{F,q}/(q\mu_q)$ is called the traffic intensity rate. A stable repair queue requires $\rho_q<1$. The repair turn-round time, denoted as t_q, measures the duration from when the part enters the repair shop to when it is fixed and returned to the spares inventory. It can be estimated by

$$t_q=\frac{B(q)}{q\mu_q-\lambda_{F,q}}+\frac{1}{\mu_q}. \qquad (6)$$

4. Parts Renewing Queueing Model

A separate Erlang-C queue, denoted as $M/M/r/\infty$, is used to characterize the part renewing process. Note that parts under renewing is a still functional item. The probability that an incoming part needs to wait in the shop prior to being renewed can be estimated as

$$C(r)=\frac{\frac{(r\rho_r)^r}{r!(1-\rho_r)}}{\sum_{j=0}^{r-1}\frac{(r\rho_r)^j}{j!}+\frac{(r\rho_r)^r}{r!(1-\rho_r)}} \qquad (7)$$

where $\rho_p=\lambda_{F,p}/(p\mu_p)$ is called the renewing traffic intensity rate with $\rho_p<1$. Note that $\lambda_{F,p}$ is the parts arrival rate to the renewing shop, and μ_p is the renewing rate per server. The renewing turn-around time, denoted as t_p, can be estimated as

$$t_r=\frac{C(r)}{r-\lambda_{F,r}}+\frac{1}{\mu_r}. \qquad (8)$$

Let $F(\tau)$ be the probability of a failure placement, and $R(\tau)$ be the probability of a proactive replacement. By combining Eqs. (6) and (8), the average part turn-around time (ATT), denoted as t_{ATT}, is obtained as follows,

$$t_{ATT}=t_qF(\tau)+t_rR(\tau)$$

$$=F(\tau)\left(\frac{B(q)}{q\mu_q-\lambda_{F,q}}+\frac{1}{\mu_q}\right)+R(\tau)\left(\frac{C(r)}{r\mu_r-\lambda_{F,r}}+\frac{1}{\mu_r}\right) \qquad (9)$$

5. The Reliability and Availability of Active Redundant System

5.1. Reliability of Single-Item System

The Reliability of a single item systems component can be determined by the well-known reliability formula used for Weibull Distribution is:

$$R(t) = exp((-\alpha * t)^{\beta}) \tag{10}$$

Where alpha is the 1/scale and beta are the shape parameter for Weibull distribution that is considered to this study. And *t denotes* as the time.

5.2. Reliability of *k*-out-of-*n* Redundant System

For a *k*-out-of-*n* redundant system with active components, the system is functional provided that at least *k* components are good at any point in time. Hence the system reliability would be:

$$\mathrm{R_{sys}}(\mathrm{k}, \mathrm{x}, \mathrm{tau}) = \sum_{j=k}^{k+x} \binom{k+x}{j} A^{j}(1-A)^{k+x-j} \tag{11}$$

Here *x* is the number of redundancies, *k* and *x* together comprised *n*.

5.3. Availability of Single-Item System

Steady-state system availability is commonly used to manage preventive maintenance and spare parts logistics when the system unitization is relatively stable (Louit et al., 2011; De Smidt-Destombes et al., 2009). It can be expressed as

$$A = \frac{MTBR}{MTBR + MDT} \tag{12}$$

where MTBR is the mean-time-between-replacements given in Eq. (1), and MDT stands for the mean downtime of a system, either due to a planned or a failure replacement. MDT under a planned replacement comprises of hands-on replacement time and delay time if the inventory is out of stock. Let *O* be a random variable representing the spare part demand of the inventory, and *s* be the base-stock level with one-for-one replenishment. The mean downtime under a planned replacement, denoted as T_p, can be expressed as

$$T_p = t_s + t_p \Pr\{O > s\}, \tag{13}$$

where t_s is the hands-on replacement time, and $\mathrm{P_r}\{O>s\}$ is the stockout probability. Similarly, the mean downtime of a failure replacement, denoted as T_q, can be expressed as

$$T_q = t_s + t_q \Pr\{O > s\}. \tag{14}$$

By combining both scenarios, the actual mean downtime of a single-item system is given as

$$\begin{aligned} MDT &= T_p R(\tau) + T_q F(\tau) \\ &= t_s + \left(t_p R(\tau) + t_q F(\tau)\right) \Pr\{O > s\} \\ &= t_s + t_{ATT} \Pr\{O > s\}. \end{aligned} \tag{15}$$

where t_{ATT} is the average part turn-around time in Eq. (9). Since O is the fleet spare parts demand that follows the Poisson process, the stockout probability of the spares inventory can be obtained as

$$\Pr\{O > s\} = \sum_{j=s+1}^{\infty} \frac{\mu^j e^{-\mu}}{j!} = 1 - \sum_{j=0}^{s} \frac{\mu^j e^{-\mu}}{j!}, \qquad \text{for } s=0, 1, 2, \ldots \tag{16}$$

with

$$\mu = \lambda_F \left(t_p R(\tau) + t_q F(\tau)\right), \tag{17}$$

where μ represents the mean demand for spare items during ATT, and λ_F is the parts demand rate. It shows that the demand during ATT depends on: (1) the demand rate, (2) the duration of ATT, and (3) the maintenance interval τ.

For a fleet comprises m single-item systems, we have $\lambda_F=\lambda_m(\tau)$ as shown in Eq. (4). Now the single-item system availability, denoted as A, is obtained by substituting Eqs. (1), (15) and (16) into (12) as follows,

$$A = \frac{\int_0^{\tau} R(t)dt}{\int_0^{\tau} R(t)dt + t_s + \left(t_p R(\tau) + t_q F(\tau)\right)\left(1 - \sum_{j=0}^{s} \mu^j e^{-\mu} (j!)^{-1}\right)}. \tag{18}$$

Note that Eq. (16) incorporates nine performance drivers. These are the component reliability $R(t)$, the maintenance interval τ, the spare parts base stock level s, the fleet size m, the hands-on replacement time t_s, the renewing and repair servers p and q, and the parts renewing rate μ_p and repair rate μ_q that are embedded in μ through Eqs. (6), (8) and (15).

5.2. Availability of *k*-out-of-*n* *Redundant System*

For a *k*-out-of-*n* redundant system with active components, the system is functional provided that at least *k* components are good at any point in time. Hence the system availability, denoted as A_R, is estimated by

$$A_R(x, s, \tau, p, q) = \sum_{j=k}^{n} \binom{n}{j} A^j (1-A)^{n-j} = \sum_{j=k}^{k+x} \binom{k+x}{j} A^j (1-A)^{k+x-j} , \tag{19}$$

where x is the number of redundant units with x+k=n. Note that A is the availability of a single-item system in Eq. (16). Together with x, there are ten performance drivers in A_R. For a fleet of m redundant systems, the spare parts demand rate is

$$\lambda_F = (x + k)\lambda_m(\tau) \tag{19b}$$

Literature Review

The optimization of system reliability and availability for automatic test equipment (ATE) has garnered significant attention, particularly in integrating redundancy allocation, maintenance scheduling, spare parts management, and repair server optimization under uncertainty. This literature review examines three key studies that address redundancy allocation problems (RAPs) and multi-objective stochastic optimization, providing insights into methodologies, contributions, and limitations relevant to the proposed two-stage stochastic programming (TSSP) model for ATE systems. These studies are categorized into two groups: redundancy allocation, which focuses on optimizing component redundancy and reliability, and multi objective/stochastic optimization, which incorporates uncertainty and multiple objectives. The reviewed works employ various failure distributions (e.g., Weibull, gamma, exponential) and series-parallel system structures, aligning with ATE's modular design. By synthesizing these studies, this review identifies gaps in maintenance, spares, and repair server considerations, highlighting the novelty of the proposed TSSP model that integrates these elements for ATE optimization.

1.1 Redundancy Allocation

1.1.1 Jin et al. (2024)

Jin et al. (2024) address the problem of minimizing the lifetime cost of ATE by jointly optimizing reliability-redundancy allocation, preventive maintenance schedules, spare parts inventory, and repair capacity under decentralized repair settings. This work is central to the current project, as it integrates redundancy, maintenance, and spares within a framework applicable to ATE, aligning with the proposed TSSP model for maximizing availability and minimizing costs. The authors formulate a redundancy-maintenance- inventory allocation (RMIA) model as a mixed-integer, nonlinear optimization problem to minimize annualized costs while meeting reliability and availability targets. Super imposed renewal theory is used to estimate spare parts demand, and parallel Erlang-C Queueing models characterize repair and renewal shops, incorporating Weibull-distributed failure times. A bisection search algorithm with neighborhood exploration solves the optimization problem, with comparisons to particle swarm optimization (PSO) and genetic algorithms (GA). The model is applied to ATE in the semiconductor industry, with Weibull-distributed component lifetimes and fleet sizes ranging from 10 to 200 systems. For a fleet of 50 systems, the optimal solution yields no redundancy ($x = 0$), 15 spares ($s = 15$), a replacement interval of 3.706 years ($\tau = 3.706$), and two servers each for repair and renewal ($p = 2$, $q = 2$), achieving 0.9901 availability at a cost of $126,407.18. For series-parallel systems with four subsystems, costs decrease with fleet size (e.g., $776,664 for $m = 10$ to $610,007 for $m = 100$), with redundancy preferred for smaller fleets or lower-cost subsystems. A key insight is that redundancy is favored for high availability requirements (e.g., 0.999), small fleets, high inventory costs, or long replacement times, with no monotonic relationship between spares and availability. The paper's holistic approach, using Weibull distributions, directly addresses gaps in ATE optimization.

1.2 Multi-Objective and Stochastic Optimization

1.2.1 Baladeh and Zio (2020)

Baladeh and Zio (2020) develop a two-stage stochastic programming (TSSP) model to optimize component test plans and redundancy allocation for series-parallel systems, maximizing reliability while minimizing variance and cost under uncertain working conditions. This study is critical to the current project, as its TSSP framework for handling uncertainty in redundancy allocation aligns with the proposed stochastic approach for ATE optimization. The TSSP model uses scenario analysis to address uncertainty, optimizing test plans and redundancy to balance expected reliability, variance, and costs via the weighted sum method (WSM). A genetic algorithm (GA) solves the multi-objective, NP-hard redundancy allocation problem, generating scenario-specific redundancy allocations. The model is applied to a feedback control system (FCS) with four subsystems (controllers, actuators, process, sensors), under three working condition scenarios with probabilities (25%, 40%, 35%). For constraints ($W_{max} = 40$, $R_{min} = 0.7$, $\mu = 0.01$, $\delta = 0.1$), the TSSP yields a stochastic solution of -0.3631, with test plans (e.g., $t12 = 5$) and redundancy achieving reliabilities of 0.8403, 0.7754, and 0.9245 across scenarios. Component reliabilities are estimated from tests with variance $v_{ij} = r_{ij} (1- r_{ij})/n_{ij}$, adaptable to the project's Weibull-based failure modeling. The staged approach, deferring redundancy decisions until working conditions are realized, enhances reliability under uncertainty, applicable to ATE's dynamic conditions. The absence of maintenance scheduling and spare parts management highlights a gap addressed by the proposed model. The FCS case study, analogous to ATE’s modular subsystems, underscores the value of scenario-based planning for the current project.

1.2.2 Jahromi and Feizabadi (2017)

Jahromi and Feizabadi (2017) propose a multi-objective redundancy allocation problem (MORAP) for series-parallel systems, maximizing reliability and minimizing cost using non-homogeneous components with mixed redundancy strategies (active and standby). This study is relevant to the current project, as it optimizes redundancy for reliability and cost, applicable to ATE subsystems, though it lacks TSSP and maintenance/spares integration. The MORAP model incorporates reliability and cost objectives, allowing non-homogeneous components in subsystems with mixed redundancy strategies, subject to weight constraints. The Non-dominated Sorting Genetic Algorithm II (NSGA-II) generates Pareto-optimal solutions for subsystems with active, standby, or mixed redundancy, handling non-repairable components. The model is tested on a series-parallel system with 14 subsystems, each with up to 4 component types, using gamma-distributed failure times with scale (λ_{ij}) and shape (k_{ij}) parameters. NSGA-II yields a Pareto front with solutions balancing reliability (e.g., $R(t) \approx 0.9$) and cost ($500–$700), with mixed strategies outperforming homogeneous designs by up to 5% in reliability for similar costs. Gamma distributions for failure times could be adapted to the project’s Weibull distributions for ATE reliability modeling. Non-homogeneous components and mixed redundancy strategies enhance flexibility in subsystem design, valuable for ATE’s modular architecture. The absence of stochastic programming and maintenance/spares integration limits its applicability to the proposed TSSP-based approach.

1.3 Synthesis and Gaps

The reviewed studies provide critical insights into redundancy allocation and multi-objective and two stage stochastic optimization for series-parallel systems, directly applicable to ATE’s modular architecture. Jin et al. (2024) offer comprehensive and ATE-specific redundancy allocation frameworks. Baladeh and Zio (2020) emphasize TSSP’s ability to handle uncertainty, aligning with the proposed stochastic framework, and Jahromi and Feizabadi (2017) highlight the benefits of non-homogeneous components. A critical gap in most studies, except Jin et al. (2024), is the absence of maintenance scheduling, spare parts management, and repair server optimization, central to ATE availability. The proposed TSSP model addresses these gaps by integrating redundancy, maintenance, spares, and repair servers, building on the strengths of these studies. The focus on ATE-specific subsystems extends general series-parallel models to industry-specific applications. These studies collectively underscore the need for a holistic approach combining stochastic programming, failure distribution modeling, and comprehensive resource allocation. The proposed work aims to bridge these gaps, enhancing ATE optimization with multi objectives through a unified TSSP framework.

The list of decision variables:

Notation	Definition	Type
x	component redundancy level (1st stage decision variable)	Integer
τ	replacement age or interval (1st stage decision variable)	Continuous
y_s	inventory stock level at scenario *s* (2nd stage decision variable)	Integer
r_s	number of renewing servers at scenario *s* (2nd stage decision variable)	Integer
q_s	number of repair servers at scenarios *s* (2nd stage decision variable)	Integer

Parameters:

Notation	Definition	Data/Source
S	Set of scenarios indexed by $s \in S$	*S*= {Optimistic, Most Likely, Pessimistic}
P_s	Probability of scenario *s*	P_s= {.2,.7,.1}
α_s, β	Weibull scale (failure rate in scenario s) and shape parameters, respectively	$\alpha_s = [0.1, 0.15, 0.2]$ $\beta = [3,4,5]$
$C_I(x,k)$	First-stage investment cost	Equation 21
C_{op}	Second-stage operational cost in scenario s	Equation 22
$R^s(x,\tau)$	Scenario specific System Reliability	Equation 24

$R^s(\tau)$	Scenario specific component's reliability function.	Equation 25
$A(x, \tau, y_s, p_s, q_s, \xi_s)$	System Availability in scenario *s*	Equation 26
A_s	Components Availability in scenario *s*	Equation 27
${\lambda^s}_m$	Aggregate part demand rate of a fleet with *m* single-item systems in scenario *s*	
${\lambda^s}_{F,p}$	Aggregate part demand rate of a fleet in planned replacement in scenario *s*	
${\lambda^s}_{F,q}$	Aggregate part demand rate of a fleet in failure replacement in scenario *s*	
${\lambda^s}_F$	Aggregate part demand rate of a fleet in scenario *s*, ${\lambda^s}_F = {\lambda^s}_{F,p} + {\lambda^s}_{F,q}$	
${\rho^s}_p, {\rho^s}_q$	Part renewing and repairing traffic intensity rate in scenario *s*	
φ_1	Capital recovery factor of system	0.1295
φ_2	Capital recovery factor of spare part	0.2310
θ	Percentage of mean time between failures of LRU (Line replaceable units)	$\theta = .7$ (approximately)
μ^s	Number of returned items during parts turn-around time in scenario *s*	

μ_p, μ_q	Parts renewing rate and repair rate respectively	$\mu_p = \frac{1}{6}$ /day $\mu_q = \frac{1}{12}$ /day
C_{LRU}	Unit cost of LRU item	$C_{LRU} = 50{,}000$ \$/item
c_h	Holding cost per item per year	$c_h = 10{,}000$ \$/part/year
c_u, c_v	Cost of renewing and repairing a part respectively	$c_u = 3{,}000$ \$/item $c_v = 4{,}500$ \$/item
c_p, c_q	Cost for operating a renewing and repairing server respectively	$c_p = 480{,}000$ \$/server $c_q = 640{,}000$ \$/server
n	Number of total LRU items	
k	The minimum number of required working items in a system	k =10 items
m	System fleet size or installed base	m = 50 units
n_{max}	Maximum number of components a system can install	n_{max} = 13 items (can vary)
$t^s{}_p$	Part renewing turn-around time	
$t^s{}_q$	Part repairing turn-around time	
t_h	Hands-on time for replacing a part	t_h= 8 hours = 0.000913 years
$B^s(q)$	Probability of a part waiting in a	

	repair queue	
$C^s(p)$	Probability for a part waiting in a renewing queue	
$\bar{T}$	Mean time between failures of the LRU	Equation 30
N	Number of redundant subsystems in a system	N=1

Assumptions:

The model relies on the following assumptions:

1. System consists of only one (N=1) k-out-of-n subsystem, with identical components within the subsystem. (This can be changed, like we can also model the system simply as a parallel system, it depends on the configuration of the system)

2. Component lifetimes follow a Weibull distribution, allowing time-varying failure rates. (This also can be changed depending on the nature of the component failure, but nowadays Weibull distribution is more popular for its versatility)

3. Repairs are perfect, restoring components to as-good-as-new condition. (This one can't be changed)

4. Decentralized repair and renewal shops follow Erlang-C queueing models ($M/M/p_s/\infty$ and $M/M/q_s/\infty$). (We can try different queuing models if the servers work differently other than exponential, like the general distribution one)

5. Age-based maintenance: proactive replacement at τ, corrective replacement upon failure. (This is fixed)

6. Spare parts inventory may experience backorders. (If the spare parts demand rates go higher than this is must, however if the demand rates are always lower than this might not be the case.

7. System utilization is stable on average. (this might not be the case all the time, therefore resulting into an infeasible solution)

8. Repair and renewal capacities are limited by server counts. (this is restricted)

9. Costs include capital, maintenance, inventory, and server operation. (We can add more cost factors depending on its necessity)

10. Queue stability requires traffic intensity rates $\rho_p, \rho_q < 1$. (this is must)

11. Components failure rate is uncertain, all other input parameters will remain the same. (we can change this assumption depending upon the problem, for example we can add uncertainty about the components cost, inventory holding cost or the cost of repair and renewal servers as well)

12. There are three possible scenarios for the component failure rate: Optimistic, Most Likely and Pessimistic. (we can consider more or less scenarios)

    I. Optimistic: Higher α, implying longer lifetimes (lower failure rate).

    II. Most Likely: Baseline α, representing typical or expected conditions.

    III. Pessimistic: Lower α, implying shorter lifetimes (higher failure rate).

The optimization method(s) applied:

This project is modeled as a Multi Objective Two-Stage Stochastic Optimization (MOTSSO) model and the problem is a Mixed Integer Nonlinear Problem (MINLP).

The type of model(s) this problem resembles:

This problem does not closely resemble any single standard Operations Research model such as the traveling salesman problem, knapsack problem, or network models (shortest path, maximum flow, minimum cost flow). It shares partial resemblance with:

- The multi-dimensional knapsack problem, due to the allocation of discrete resources (redundancy x, inventory y, servers p, q) to optimize objectives under constraints like availability and physical limits.
- The minimum cost flow problem, as the spare parts logistics and repair/renewal processes can be viewed as flows with server capacities, minimizing costs while meeting demand.

However, the multi-objective nature (cost vs. reliability/availability), two-stage stochastic framework (with failure rate scenarios), nonlinear constraints (availability, queueing), and integration of reliability, maintenance, and inventory decisions make this problem unique and hybrid as well. It extends beyond classical models, resembling a generalized stochastic multi-objective resource allocation problem tailored to reliability and logistics systems.

The mathematical formulation:

The problem can be formulated as a Multi-Objective Two Stage Stochastic Optimization. The 1st stage makes strategic decision on redundancy level x, and the maintenance interval tau (τ), where the objective is minimizing the total investment cost and maximizing the system reliability. The 2nd stage decision is associated with the system operational period, and it involves repair and renewal of components as well as the spare parts provisioning. Component reliability or component failure time is the preliminary uncertain factor that influences the spares parts demand and system availability. Therefore, spare parts inventory y_s, the number of renewal servers p_s, and the number of repair server q_s are the decision variables during the operational phase. The goal is to reduce the annualized system cost while meeting the minimum system operational availability.

Objective 1: Minimize Total Cost

$$\min_{x,\tau,y_s,p_s,q_s} C_I(x,\tau) + \sum_{s\in S} P_s \cdot C_{op}(x,\tau,y_s,p_s,q_s,\xi_s) \tag{20}$$

Here we are minimizing the investment cost and the expected cost for the operational phase.
The first stage investment cost is defined as:

$$C_I(x,k) = (k+x)\cdot(\varphi_1 \cdot C_{\mathrm{LRU}}) \tag{21}$$

Where the operational cost of the second stage is like this:

$$C_{op}(x,\tau,y_s,p_s,q_s,\xi_s) = (k+x)\left(\lambda_p^s c_u + \lambda_q^s c_v\right) + \frac{1}{m}\left(y_s(\varphi_2 C_{\mathrm{LRU}} + c_h) + \left(p_s c_p + q_s c_q\right)\right) \tag{22}$$

Objective 2: Maximize Expected System Reliability

$$\max_{x,k,\tau} \sum_{s\in S} P_s\, R^s(x,\tau) \tag{23}$$

$$R^s(x,\tau) = \binom{k+x}{j} R(\tau)^j \left(1 - R(\tau)\right)^{k+x-j}$$

(24)

$$R^s(\tau) = \exp(-\alpha_s \tau)^{\beta} \tag{25}$$

Subject to:

$$A(x, \tau, y_s, p_s, q_s, \alpha_s) \geq A_{min}$$

(26)

Here our constraint is to meet the minimum availability of the system that is required for each scenario. Now the availability of the system can be found by the following formula:

$$A(x, \tau, y_s, p_s, q_s, \alpha_s) = \sum_{j=k}^{n} \binom{n}{j} A_s{}^{j} (1 - A_s)^{n-j} = \sum_{j=k}^{k+x} \binom{k+x}{j} A_s{}^{j} (1 - A_s)^{k+x-j} \tag{27}$$

Whereas the scenario specific components' availability can be defined as:

$$A_s = \frac{\int_0^{\tau} R^s\,(t)\,dt}{\int_0^{\tau} R^s\,(t)\,dt + t_h + \left(t_p^s R^s(\tau) + t_q^s F^s(\tau)\right)\left(1 - \sum_{j=0}^{y_s} \frac{\mu_s^j e^{-\mu_s}}{j!}\right)^{-1}} \tag{28}$$

Now we introduce the maintenance time interval constraint, we cannot do regular maintenance frequently as it involves cost and over maintenance as well, therefore we required a lower bound for the maintenance interval which is:

$$\tau \geq \theta\,\overline{T} \tag{29}$$

Where $\overline{T}$ is the mean time between failure and theta represent the percentage of it that we are putting as the lower bound for the interval.

$$\overline{T} = \frac{1}{\alpha} \times \Gamma\left(1 + \frac{1}{\beta}\right) \tag{30}$$

Now we cannot have as many redundant units as we want, because the system has a maximum capacity as well. Therefore:

$$x + k \leq n_{\max} \tag{31}$$

The traffic intensity rate of the repair and renewal shop must be less than 1 for a stable queue.

$$\rho_p^s < 1 \tag{32}$$

$$\rho_q^s < 1 \tag{33}$$

The redundant units must be the nonnegative integer, and the time interval is a nonnegative continuous decision variable.

$$x \in \mathbb{Z}_{\geq 0}, \quad \tau \geq 0 \tag{34}$$

$$y_s \in \mathbb{Z}_{\geq 0}, \quad p_s \in \mathbb{Z}^+, \quad q_s \in \mathbb{Z}^+, \quad \forall s \in S \tag{35}$$

Scalarized Final Objective Function:

This model will be solved using the weighted sum method. Let w_1 and w_2 be the respective weights for total cost minimization and reliability maximization, Then the final objective function will be:

$$\min_{x,\tau,y_s,p_s,q_s} w_1 \cdot \left(C_x(x,\tau) + \sum_{s \in S} P_s \cdot C_{op}(x,\tau,y_s,p_s,q_s,\alpha_s) \right) - w_2 \cdot \left(\sum_{s \in S} P_s \, R(x,\tau) + \right) \tag{36}$$

Subject to:

$$A(x,\tau,y_s,p_s,q_s,\alpha_s) \geq A_{min} \tag{37}$$

$$\tau \geq \theta \, \overline{T} \tag{38}$$

$$x + k \leq n_{\max} \tag{39}$$

$$x \in \mathbb{Z}_{\geq 0}, \quad \tau \geq 0 \tag{40}$$

$$\rho_r^s < 1 \tag{41}$$

$$\rho_q^s < 1 \tag{42}$$

$$y_s \in \mathbb{Z}_{\geq 0}, \quad p_s \in \mathbb{Z}^+, \quad q_s \in \mathbb{Z}^+, \quad \forall s \in S \tag{43}$$

Numerical results and discussion

This model is solved using Gurobi version 12.0.1 with a weighted sum method, where I use w1=1 and w2=1000000 to balance between the cost and the reliability.

(a) The number of decision variables and constraints

Decision Variables: The model has 483 columns, with 276 continuous and 207 integer variables (197 binary). Thus, the total number of decision variables is 483.

Constraints: The model starts with 58 rows (linear constraints), 18 quadratic constraints, 414 simple general constraints (207 indicator, 207 PWL), and 21 general nonlinear constraints (24 nonlinear terms). After presolve, there are 1342 rows, including 207 SOS constraints, 45 bilinear constraints, and 9 nonlinear constraints. The initial number of constraints is 58 (linear) + 18 (quadratic) + 414 (general) + 21 (nonlinear) = 511 constraints, excluding SOS and bilinear constraints added during presolve.

(b) The optimal values for decision variables

The optimal solution of the Multi Objective Two Stage Stochastic MINLP models with the baseline data are:

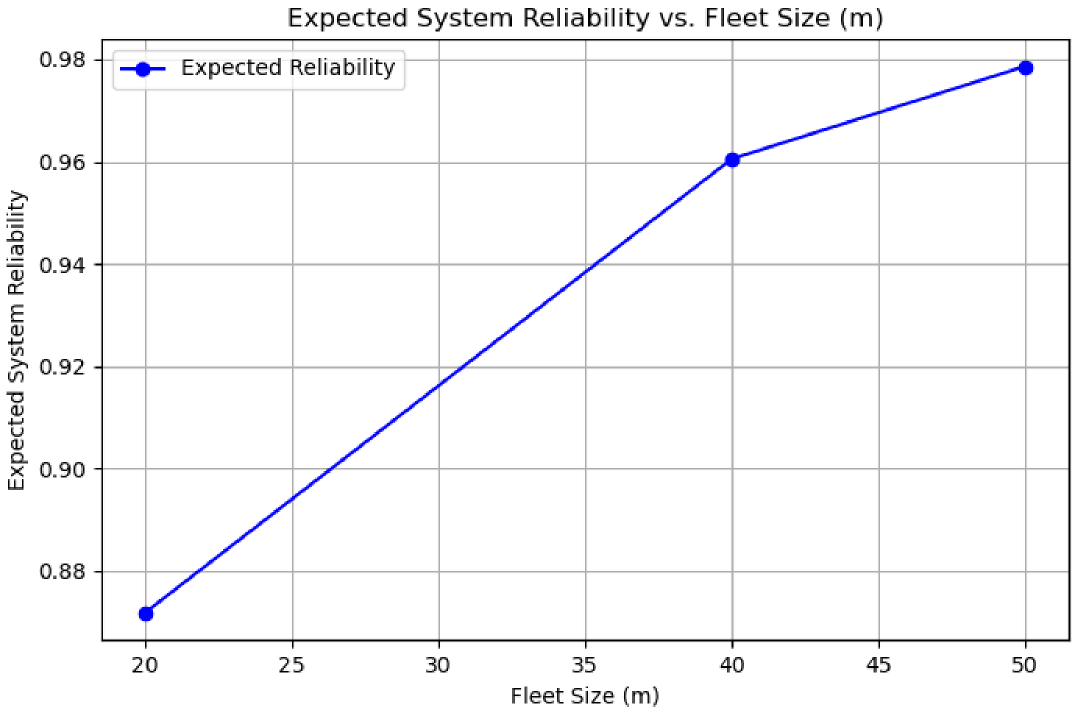
- *x* = 3
- tau = 2.7778
- Optimistic: *r*=4, *q*=1, *y*=0
- Most Likely: *r*=4, *q*=1, *y*=0
- Pessimistic: *r*=4, *q*=2, *y*=0

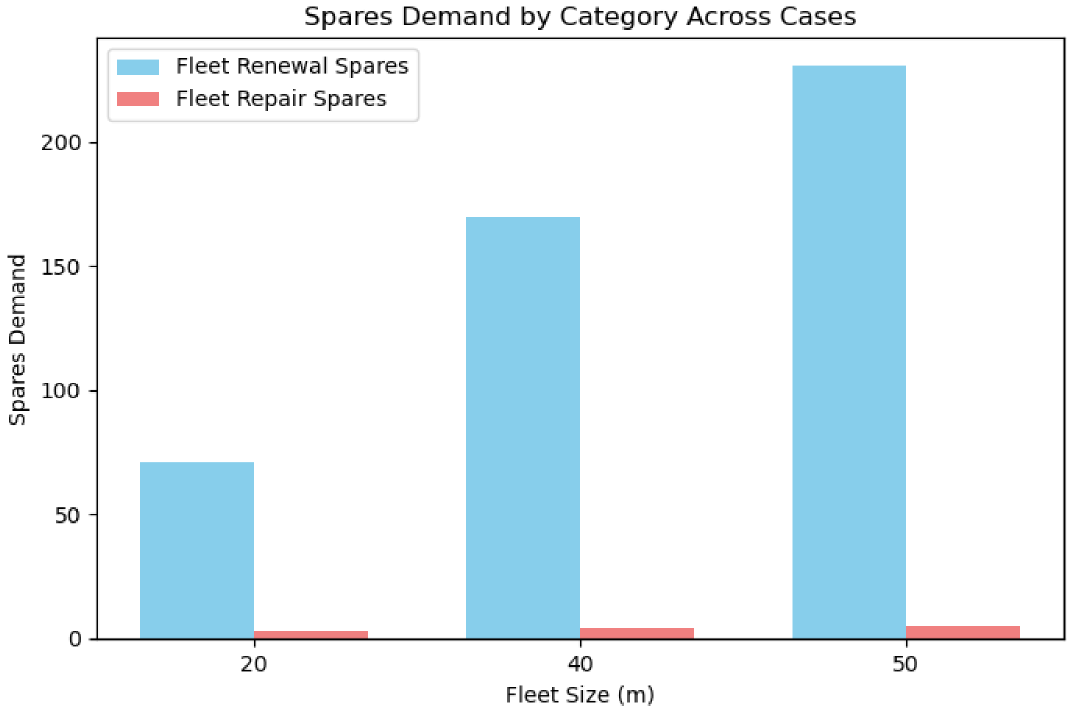
(c) Sensitivity analysis.

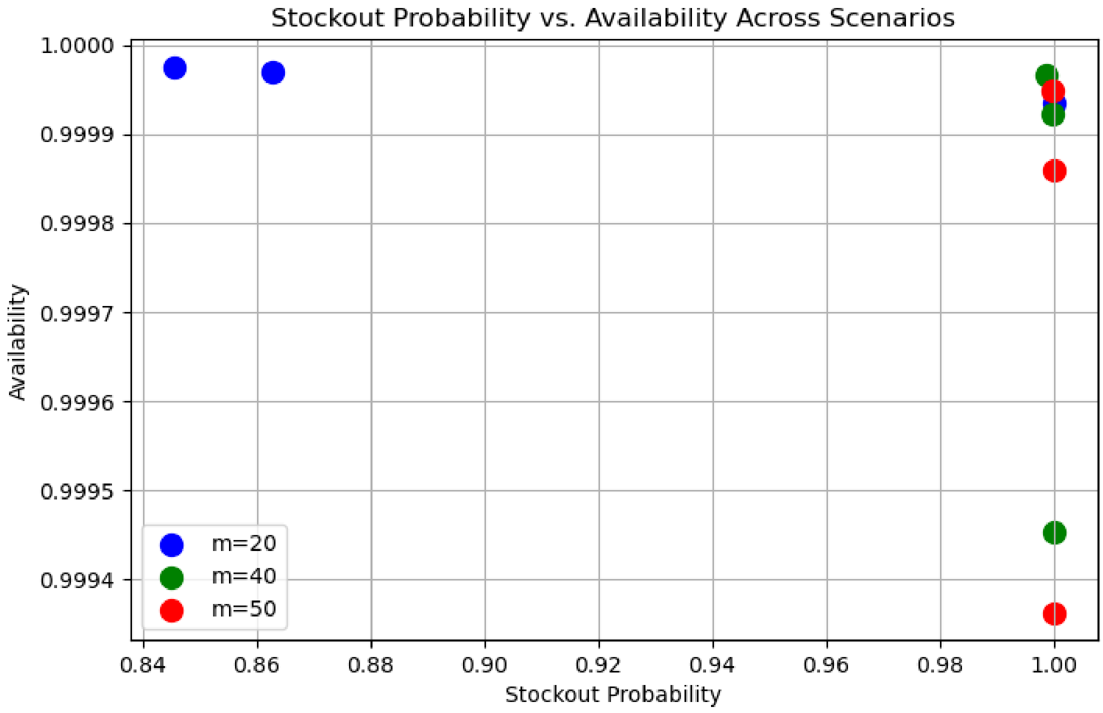
Table1: Sensitivity Analysis for three cases of fleet size

Metric	Case 1 (m=20)	Case 2 (m=40)	Baseline (m=50)
Optimal Expected Cost	173,945.08	151,642.07	151,441.56
x (Decision Variable)	3	3	3
tau	3.5556	3	2.7778
Expected System Reliability	0.8719	0.9605	0.9786

Optimistic Stockout Prob	0.8455	1	1
Optimistic Availability	0.999976	0.999454	0.999363
Optimistic Reliability	0.9978	0.9995	0.9998
Most Likely Stockout Prob	0.8626	0.9999	1
Most Likely Availability	0.999969	0.999923	0.999859
Most Likely Reliability	0.9013	0.9775	0.989
Pessimistic Stockout Prob	1	0.9987	0.9998
Pessimistic Availability	0.999935	0.999967	0.999949
Pessimistic Reliability	0.4146	0.7641	0.8636
Fleet Renewal Spares Demand	70.695511	169.85414	230.26476
Fleet Repair Spares Demand	3.25021	4.648535	4.986644


Expected System Reliability vs. Fleet Size (m)
Expected Reliability
Expected System Reliability
0.98
0.96
0.94
0.92
0.90
0.88
20
25
30
35
40
45
50
Fleet Size (m)


Spares Demand by Category Across Cases
Fleet Renewal Spares
Fleet Repair Spares
Spares Demand
200
150
100
50
0
20
40
50
Fleet Size (m)



The sensitivity analysis evaluates the impact of fleet size (*m*) on system performance metrics, including cost, reliability, availability, and spares demand. As m increases from 20 to 50:

- Expected Cost decreases (from 17,3945.08 to 15,1441.56), suggesting economies of scale or optimization efficiency.
- Expected System Reliability improves (from 0.8719 to 0.9786), indicating larger fleets enhance system robustness.
- Spares Demand for Fleet Renewal rises sharply (from 70.7 to 230.3), reflecting higher maintenance needs with larger fleets, while Fleet Repair Demand remains low and stable.
- Stockout Probability approaches 1 in optimistic and most likely scenarios, suggesting low risk, but Pessimistic Reliability varies widely (0.4146 to 0.8636), indicating sensitivity to worst-case conditions.

This analysis suggests that increasing m improves reliability and cost efficiency but significantly increases renewal spares demand, which may require strategic inventory planning.

Table2: Sensitivity Analysis for three cases of k and tau combination

Metric	**Case 4 (k=5, tau=2.7778)**	**Case 5 (k=5, tau=3.0000)**	**Case 6 (k=10, tau=3.0000)**

Optimal Expected Cost	192499.42	190973.58	158963.76
x (Decision Variable)	3	3	4
tau	2.7778	3	3
Expected System Reliability	0.9964	0.9926	0.9842
Optimistic Stockout Prob	0.9895	0.972	1
Optimistic Availability	1	1	0.999962
Optimistic Reliability	1	0.9999	0.9999
Most Likely Stockout Prob	0.9873	0.9732	1
Most Likely Availability	1	1	0.999997
Most Likely Reliability	0.9985	0.9968	0.9945
Pessimistic Stockout Prob	0.9958	1	0.9998
Pessimistic Availability	1	0.999867	0.999997

Pessimistic Reliability	0.9744	0.9484	0.8812
Fleet Renewal Spares Demand	141.70139	130.65703	228.64988
Fleet Repair Spares Demand	3.699935	3.575796	6.257643

This sensitivity analysis for Case 4, Case 5, and Case 6 focuses on the impact of varying k and tau on system performance:

- Expected Cost is lowest in Case 6 (*k*=10, tau=3.0000) at 158963.76, compared to 192499.42 (Case 4) and 190973.58 (Case 5), indicating that increasing k reduces costs significantly.
- Expected System Reliability is highest in Case 4 (0.9964) with *k*=5 and tau=2.7778, and lowest in Case 6 (0.9842) with *k*=10, suggesting that a lower tau improves reliability for smaller *k*.
- Spares Demand for Fleet Renewal spikes in Case 6 (228.64988) due to the higher *k*, while Case 5 has the lowest demand (130.65703). Repair spares demand remains low but increases slightly with *k*.
- Stockout Probability and Availability remain near 1 across most scenarios, but Pessimistic Reliability drops to 0.8812 in Case 6, indicating sensitivity to worst-case scenarios with higher *k*.

Table3: Sensitivity Analysis for reduced weight for w2

Metric	**Case 8 (k=10, tau=2.5000)**	**Case 9 (k=10, tau=2.7778)**	**Case 10 (k=10, tau=2.7778)**
w1	1	1	1
w2	100000	100000	100000
Optimistic (r, q, y)	(4, 1, 0)	(4, 1, 9)	(4, 1, 0)
Most Likely (r, q, y)	(4, 1, 0)	(4, 1, 3)	(4, 1, 0)
Pessimistic (r, q, y)	(4, 1, 0)	(4, 2, 0)	(4, 2, 0)
Optimal Expected Cost	136138.97	67352.41	143829.13
x (Decision Variable)	1	3	2

tau	2.5	2.7778	2.7778
Expected System Reliability	0.8833	0.9786	0.9369
Optimistic Reliability	0.9846	0.9998	0.998
Most Likely Reliability	0.8915	0.989	0.952
Pessimistic Reliability	0.6234	0.8636	0.7092

Reducing w2 from 1000000 to 100000 decreases the focus on avoiding stockouts, leading to lower reliability (e.g., Case 8’s 0.8833 vs. Case 9’s 0.9786) because the system allocates fewer resources (*x*) or less optimal tau. Case 9’s higher *q* and *x* mitigate this effect, achieving the best reliability.

(d) The results make practical sense

Yes, the results make practical sense. The total expected cost (151441.56) and expected system reliability (0.9786) align with the optimization’s objectives, balancing cost and reliability under the given weights (*w*1=1, *w*2=1000000). The decision variables (*x*=3, tau=2.7778) lead to high reliability (0.9998 optimistic, 0.9890 most likely) and reasonable spares demand (e.g., 230.264763 for fleet renewal), reflecting realistic resource allocation. The pessimistic reliability (0.8636) with *q*=2 indicates sensitivity to worst-case scenarios, which is practical for a non-convex MINLP model with trade-offs between cost, reliability, and spares demand. The model’s behavior matches expectations for a system prioritizing reliability while managing costs.

(e) Comparison with different solutions

The Gurobi solution (Case 9: *x*=3, tau=2.7778, cost=151441.56, reliability=0.9786) aligns with Case 3 (*m*=50), showing consistency. Compared to Cases 1-10, Case 9 balances cost and reliability well, outperforming Cases 4 and 5 (higher cost: 192499.42, 190973.58; higher reliability: 0.9964, 0.9926) and Case 7 (high cost: 950745.08; low reliability: 0.8719). Higher *x* and moderate tau in Case 9 optimize reliability, while *w*2=100000 in Cases 8-10 lowers costs but reduces reliability when *x* is low (Case 8: 0.8833). Varying *p*, *q* in Case 9 enhances resilience, making the results practical for fleet management trade-offs.

Conclusion

This study presents a comprehensive approach to optimizing redundancy, maintenance, spare parts, and repair server allocation under uncertainty to minimize costs while maximizing system reliability and availability. By developing a robust system availability model that integrates ten key performance factors, we establish a foundation for a joint redundancy-inventory-repair allocation model using two-stage stochastic programming. The application of this model in the automatic test equipment (ATE) industry demonstrates its effectiveness in enhancing operational availability and reliability—critical factors for maintaining global competitiveness. Numerical analysis provides actionable managerial insights, helping decision-makers balance cost efficiency with system performance in uncertain environments. Future research could explore dynamic optimization techniques, additional sources of uncertainty, or industry-specific extensions to further refine the model's applicability.

Recommendations

Future work could extend this Multi Objective Two Stage Stochastic redundancy-inventory-repair model by:

1. Incorporating predictive maintenance with real-time sensor data
2. Expanding to multi-stage stochastic programming for dynamic uncertainties
3. Adding sustainability objectives like energy efficiency
4. Validating in other high-tech industries beyond ATE
5. Investigating human factors in maintenance operations

These extensions would enhance the model's applicability while preserving its core strengths in availability optimization and cost-effective resource allocation.

Acknowledgement

I would like to express my deepest gratitude to Professor Dr. Clara Novoa for her invaluable guidance and support throughout this semester. Her expertise in operations research, optimization techniques, two-stage stochastic programming and multi-objective optimization has been instrumental in shaping my understanding and this project. Her insightful feedback and encouragement have greatly enhanced both my technical skills and research perspective.

I am also sincerely thankful to Professor Dr. Tongdan Jin for introducing me to this project and sharing his profound knowledge of reliability, availability, and queuing models, which formed the foundation of this study.

Finally, I appreciate all the support from my colleagues in completing this work.

References

- A. E. Baladeh and E. Zio, "A Two-Stage Stochastic Programming Model of Component Test Plan and Redundancy Allocation for System Reliability Optimization," in *IEEE Transactions on Reliability*, vol. 70, no. 1, pp. 99-109, March 2021, https://doi.org/10.1109/TR.2020.2974284
- Cox, D. R., & Smith, W. L. (1954). On the superposition of renewal processes. *Biometrika,* 41, 91-99. https://doi/10.1093/biomet/41.1-2.91.

- De Smidt-Destombes, K. S., van der Heijden, M. C., & van Harten, A. (2009). Joint optimisation of spare part inventory, maintenance frequency and repair capacity for *k*-out-of-*N* systems. *International Journal of Production Economics,* 118, 260-268. https://doi.org/10.1016/j.ijpe.2008.08.058.
- El-Ferik, S. (2008). Economic production lot-sizing for an unreliable machine under imperfect age-based maintenance policy. *European Journal of Operational Research,* 186, 150-163. https://doi.org/10.1016/j.ejor.2007.01.035.
- Eshraghniaye Jahromi, A., & Feizabadi, M. (2017). Optimization of multi-objective redundancy allocation problem with non-homogeneous components. *Computers & Industrial Engineering*, *108*, 111-123. https://doi.org/10.1016/j.cie.2017.04.009
- Hekimoğlu, M., van der Laan, E., & Dekker, R. (2018). Markov-modulated analysis of a spare parts system with random lead times and disruption risks. *European Journal of Operational Research,* 269, 909-922. https://doi.org/10.1016/j.ejor.2018.02.040.
- Huynh, K. T., Castro, I. T., Barros, A., & Bérenguer, C. (2012). Modeling age-based maintenance strategies with minimal repairs for systems subject to competing failure modes due to degradation and shocks. *European Journal of Operational Research,* 218, 140-151. https://doi.org/10.1016/j.ejor.2011.10.025.
- Jin, T., Si, S. & Zhu, W. Allocating redundancy, maintenance and spare parts for minimizing system cost under decentralized repairs. *Front. Eng. Manag.* **11**, 377–395 (2024). https://doi.org/10.1007/s42524-024-0145-3
- Jin, T., Tian, Z., & Xie, M. (2015). A game-theoretical approach for optimizing maintenance, spares and service capacity in performance contracting. *International Journal of Production Economics,* 31, 31-43. https://doi.org/10.1016/j.ijpe.2014.11.010.
- Louit, D., Pascual, R., Banjevic, D., & Jardine, A. K. S. (2011). Optimization models for critical spare parts inventories-a reliability approach. *Journal of the Operational Research Society,* 62, 992-1004. https://doi.org/10.1057/jors.2010.49.
- Winston, W. (2004). *Operations Research: Applications and Algorithms* (4th ed.) (Chapter 20). Brooke/Cole Cengage Learning, Belmont, CA, USA.