ISSN 1854-6250

APEM journal

Advances in Production Engineering & Management

Volume 20 | Number 1 | March 2025

University of Maribor

Published by CPE
apem-journal.org

Advances in Production Engineering & Management

Identification Statement

ISSN 1854-6250 | Abbreviated key title: **Adv produc engineer manag** | Start year: 2006
ISSN 1855-6531 (on-line)

Published quarterly by **Chair of Production Engineering (CPE)**, **University of Maribor**
Smetanova ulica 17, SI – 2000 Maribor, Slovenia, European Union (EU)

Phone: 00386 2 2207522, Fax: 00386 2 2207990
Language of text: English

APEM homepage: apem-journal.org
University homepage: www.um.si

APEM Editorial

OPEN ACCESS

Subsidizer: The journal is subsidized by Slovenian Research and Innovation Agency

Published by CPE, University of Maribor.

Advances in Production Engineering & Management is indexed and abstracted in the **WEB OF SCIENCE** (maintained by **Clarivate**): **Science Citation Index Expanded**, **Journal Citation Reports** – Science Edition, **Current Contents** – Engineering, Computing and Technology • **Scopus** (maintained by **Elsevier**) • **Inspec** • **EBSCO**: Academic Search Alumni Edition, Academic Search Complete, Academic Search Elite, Academic Search Premier, Engineering Source, Sales & Marketing Source, TOC Premier • **ProQuest**: CSA Engineering Research Database – Cambridge Scientific Abstracts, Materials Business File, Materials Research Database, Mechanical & Transportation Engineering Abstracts, ProQuest SciTech Collection • **TEMA** (**DOMA**) • The journal is listed in **Ulrich's** Periodicals Directory and **Cabell's** Directory

University of Maribor
Chair of Production Engineering (CPE)

Advances in Production Engineering & Management

Volume 20 | Number 1 | March 2025 | pp 1–152

Contents

Journal homepage: apem-journal.org

ISSN 1854-6250 (print)
ISSN 1855-6531 (on-line)

Published by CPE, University of Maribor.

Scope and topics

Advances in Production Engineering & Management (*APEM journal*) is an interdisciplinary refereed international academic journal published quarterly by the *Chair of Production Engineering* at the *University of Maribor*. The main goal of the *APEM journal* is to present original, high quality, theoretical and application-oriented research developments in all areas of production engineering and production management to a broad audience of academics and practitioners. In order to bridge the gap between theory and practice, applications based on advanced theory and case studies are particularly welcome. For theoretical papers, their originality and research contributions are the main factors in the evaluation process. General approaches, formalisms, algorithms or techniques should be illustrated with significant applications that demonstrate their applicability to real-world problems. Although the *APEM journal* main goal is to publish original research papers, review articles and professional papers are occasionally published.

Fields of interest include, but are not limited to:

Additive Manufacturing Processes
Advanced Production Technologies
Artificial Intelligence in Production
Assembly Systems
Automation
Big Data in Production
Block Chain in Manufacturing
Computer-Integrated Manufacturing
Cutting and Forming Processes
Decision Support Systems
Deep Learning in Manufacturing
Discrete Systems and Methodology
e-Manufacturing
Evolutionary Computation in Production
Fuzzy Systems
Human Factor Engineering, Ergonomics
Industrial Engineering
Industrial Processes
Industrial Robotics
Intelligent Manufacturing Systems
Joining Processes
Knowledge Management
Logistics in Production
Machine Learning in Production
Machine-to-Machine Economy
Machine Tools
Machining Systems
Manufacturing Systems
Materials Science, Multidisciplinary
Mechanical Engineering
Mechatronics
Metrology in Production
Modelling and Simulation
Numerical Techniques
Operations Research
Operations Planning, Scheduling and Control
Optimisation Techniques
Project Management
Quality Management
Risk and Uncertainty
Self-Organizing Systems
Smart Manufacturing
Statistical Methods
Supply Chain Management
Virtual Reality in Production

Advances in Production Engineering & Management **ISSN 1854-6250**
Volume 20 | Number 1 | March 2025 | pp 5–17 Journal home: apem-journal.org
https://doi.org/10.14743/apem2025.1.523 Original scientific paper

Reinforcement learning for robot manipulation tasks in human-robot collaboration using the CQL/SAC algorithms

Husaković, A.[a]**, Banjanović-Mehmedović, L.**[b]**, Gurdić-Ribić, A.**[c]**, Prljača, N.**[b]**, Karabegović, I.**[d]

[a]Eacon doo, Zenica, Bosnia and Herzegovina
[b]University of Tuzla, Faculty of Electrical Engineering, Tuzla, Bosnia and Herzegovina
[c]Foundation for Innovation, Technology and Transfer of Knowledge, Tuzla, Bosnia and Herzegovina
[d]Academy of Sciences and Arts of Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina

ABSTRACT

The integration of human-robot collaboration (HRC) into industrial and service environments demands efficient and adaptive robotic systems capable of executing diverse tasks, including pick-and-place operations. This paper investigates the application of Soft Actor-Critic (SAC) and Conservative Q-Learning (CQL)—two deep reinforcement learning (DRL) algorithms—for the learning and optimization of pick-and-place actions within HRC scenarios. By leveraging SAC's capability to balance exploration and exploitation, the robot autonomously learns to perform pick-and-place tasks while adapting to dynamic environments and human interactions. Moreover, the integration of CQL ensures more stable learning by mitigating Q-value overestimation, which proves particularly advantageous in offline and suboptimal data scenarios. The combined use of CQL and SAC enhances policy robustness, facilitating safer and more efficient decision-making in continually evolving environments. The proposed framework combines simulation-based training with transfer learning techniques, enabling seamless deployment in real-world environments. The critical challenge of trajectory completion is addressed through a meticulously designed reward function that promotes efficiency, precision, and safety. Experimental validation demonstrates a 100 % success rate in simulation and an 80 % success rate on real hardware, confirming the practical viability of the proposed model. This work underscores the pivotal role of DRL in enhancing the functionality of collaborative robotic systems, illustrating its applicability across a range of industrial environments.

ARTICLE INFO

Keywords:
Human-robot collaboration;
Robot learning;
Deep reinforcement learning;
Soft actor-critic algorithm (SAC);
Conservative Q-learning (CQL);
Robot manipulation tasks

**Corresponding author:*
lejla.banjanovic-mehmedovic@fet.ba
(Banjanović-Mehmedović, L.)

Article history:
Received 10 February 2025
Revised 2 March 2025
Accepted 7 March 2025

1. Introduction

Human-robot collaboration (HRC), driven by Industry 4.0 and 5.0, has become a cornerstone of modern industrial and service robotics, where robots and humans work together to achieve common goals. The human worker is tasked with solving social challenges and making decisions, while the cobot dictates the speed and acceleration of the process [1]. These collaborative environments require robots that are not only efficient and precise but also adaptable to dynamic scenarios involving human interaction [2]. HRC systems, within the framework of Industry 4.0, contribute to improving workflow, accelerating processes, enhancing product quality, reducing energy consumption and CO2 emissions. In the context of Industry 5.0, HRC is worker-centered, focusing on creating a positive and conflict-free working environment [3].

In the evolution of smart robotic manufacturing, the ability to adapt to uncertainties, changing environments, and dynamic tasks has become a critical necessity, often described as the emergence of *program-free robots* or *learning-enabled robots* [4]. Robot learning integrates various machine learning techniques into robotics [5], focusing on enabling robots to learn and perform actions based on environmental inputs. This paradigm addresses challenges such as optimize policies for real-time decision-making, managing the exploration-exploitation trade-off, incorporate feedback to adapt to evolving tasks and conditions, and ensuring robust transfer of learning from simulations to real-world applications. These factors make robot learning a unique and complex area within machine learning and robotics.

Robotic manipulation is a fundamental challenge in robotics, involving tasks such as grasping, object handling, and assembly. These tasks are inherently complex due to high-dimensional state and action spaces, environmental uncertainties, and the dynamic interactions between the robot, objects, and human collaborators. Incorporating human-cobot collaboration adds another layer of complexity, as robots must effectively and safely interact with humans in shared workspaces. The safety distance between humans and machines, based on real-time changes in various scenarios, ensures smoother and safer collaboration [6]. The paper [7] emphasizes the importance of evaluating collaborative workplaces in both simulation and real-world environments, showing that such evaluations can enhance overall industrial system capacity This requires a high level of adaptability, precision, and real-time decision-making to ensure seamless task sharing, coordination, and safety in dynamic industrial or service environments [8-10].

Learning for robot manipulation is essential for enabling robots to adapt to complex, unstructured environments and perform tasks that require interacting with diverse objects [11]. Reinforcement learning (RL) has proven effective in training robots to perform intricate manipulation and grasping tasks, assembly and disassembly tasks, which is vital for advancing human-robot collaboration. By equipping robots with the ability to adapt and respond dynamically to their environment and human partners, RL enhances their role in shared workspaces across manufacturing, warehousing, healthcare, and service robotics applications [12].

Applying DRL in continual, dynamic environments is crucial for enabling robots to adapt to real-world uncertainties, optimize decision-making over time, and improve efficiency in tasks such as robotic manipulation and autonomous navigation [8]. Conservative Q-Learning (CQL) and Soft Actor-Critic (SAC) are both off-policy reinforcement learning algorithms tailored for continuous control, with SAC optimizing a stochastic policy using entropy regularization to encourage exploration, while CQL focuses on learning conservative Q-values to mitigate overestimation and improve robustness in offline learning [13]. While SAC excels in dynamic environments with abundant data, CQL is particularly useful in settings with limited or suboptimal datasets by constraining the learned Q-values to avoid risky actions [14].

Hierarchical Reinforcement Learning (HRL) is an AI approach that structures decision-making by breaking complex tasks into manageable sub-tasks, enhancing learning efficiency [15]. A Hierarchical Reinforcement Learning (HRL) approach enhances human-robot interactions by breaking complex tasks into smaller sub-tasks, allowing the robot to independently optimize each one while contributing to the overall goal. HRL improves learning efficiency by enabling the robot to solve simpler problems faster, adapts better to human inputs like gestures or commands, and offers scalability as new sub-tasks can be added without disrupting the system. It also decomposes complex tasks, such as assembly or assistance in dynamic environments, into manageable components, improving overall performance and flexibility.

This paper focuses on applying the CQL/SAC algorithm to develop a framework for learning and executing two sub-tasks, pick-and-place actions in human-robot collaborative settings. The proposed approach addresses several critical challenges, including trajectory optimization, and autonomously and safe interaction with human collaborators. By integrating reinforcement learning into the HRC domain, this paper aims to improve the robot's ability to adapt to varying conditions while maintaining task efficiency and safety.

The paper is structured as follows. Section 2 reviews related work on robot manipulation, with focus on pick-and-place operations and reinforcement learning. Section 3 presents the proposed deep reinforcement methodology, detailing the CQL/SAC-based learning framework. Section 4 discusses the system description, including the collaborative robotics platform and reward function design. Results are presented in Section 5, demonstrating the effectiveness of the CQL/SAC algorithm in optimizing pick-and-place actions. Finally, Section 6 presents the conclusion of this study and discusses potential directions for future research.

2. Related work

The precise object grasping is essential for robots, especially in industrial settings. Most learning-driven grasping tasks rely on vision signals, requiring models with strong representational capabilities. Using deep convolutional neural networks and a guided policy search method, the study presented in paper [16] enables robots to learn policies that map raw visual inputs directly to motor commands. The approach is validated on real-world tasks requiring vision-control coordination, such as assembling a bottle cap.

Reinforcement Learning (RL) enables robots to learn manipulation and collaboration skills through interaction with their environment and human partners. By optimizing a reward function, robots can develop policies to perform complex tasks and adapt to human inputs. The deep reinforcement learning and deep neural networks like CNNs are commonly used, showcasing the power of robot learning-based grasping. Mahler *et al*. used CNN and DQN for pick-and-transport tasks with an ABB Yumi robot [12]. Mohammed *et al*. utilized DQN and Q-learning with RGBD input for target position generation on UR robots [17].

Recent research has demonstrated the effectiveness of DRL in various task-specific applications [18]. Liu *et al*. [19] designed a robotic pick-and-place system, emphasizing reward shaping to address complex challenges. Rewards were based on the Euclidean distance between an object's current and target configurations, with additional rewards for task completion, significantly improving convergence compared to linear reward methods. The study also demonstrated that the employed learning technique outperformed both PPO and Actor-Critic (A3C) by directly generating manipulation signals, thereby effectively managing high task complexity. A simulated pick-and-place task with a simple block, using DDPG enhanced with hindsight experience replay (HER), is presented in [20].

In [21], a flexible framework was proposed that integrates motion planning with RL for continuous robot control in cluttered environments. By combining model-free RL with a sampling-based motion planner, this approach minimizes dependency on task-specific knowledge and enables the RL policy to determine when to plan or execute direct actions through reward maximization. Additionally, [22] introduced a framework that leverages demonstrations, unsupervised learning, and RL to efficiently learn complex tasks using only image input.

SAC algorithm, has demonstrated broad applicability in tasks like door opening and block stacking by breaking the task into fundamental steps (e.g., reaching, grasping, turning, and pulling for door opening) and training each step individually [23]. The advantage of SAC in path planning lies in entropy maximization, which encourages the robot to explore its surroundings, while the addition of HER (hindsight experience replay) allows the use of past experiences for improved learning and environmental adaptation [24]. In paper [25], a novel framework for sequentially learning vision-based robotic manipulation tasks in offline reinforcement learning settings was presented. Their approach leverages offline data and visual inputs to train adaptable policies for complex, sequential manipulation scenarios, demonstrating improved performance and generalization in robotic applications.

Multi-Agent Deep Reinforcement Learning (MADRL) involves multiple agents learning and interacting in the same environment, either through collaboration or competition. Each agent learns to optimize its own policy while considering the actions of others. It is widely used in robotics and industrial automation for tasks such as coordinated robot control, resource allocation, and adaptive decision-making in complex systems. The paper [26] employs a multi-agent deep reinforcement learning (DRL) approach, which is a significant advancement in the field of production scheduling. This method incorporates an attention mechanism within an advantage actor-critic framework, which is further complemented by a global reward function.

3. Deep reinforcement learning

Deep Reinforcement Learning (DRL) has emerged as a transformative approach, providing adaptive and intelligent solutions for optimizing robotics and production processes in complex industrial environments. It enhances efficiency by enabling autonomous systems to learn optimal strategies for navigation, manipulation, and workflow automation [27].

Finding the optimal policy in reinforcement learning is influenced by whether the method is model-based or model-free. Model-based approaches estimate transition probabilities p(s′|s,a) using prior knowledge or state-space searches, while model-free methods learn directly from rewards without modelling system dynamics. Model-free techniques are common in robotics due to the complexity of modelling continuous state spaces.

RL requires balancing exploration (random actions to discover optimal strategies) and exploitation (choosing actions to maximize rewards). This trade-off is managed using a parameter ε, which controls the likelihood of exploring non-optimal actions. On-policy learning adjusts the current policy by mixing optimal and random actions, while off-policy learning uses a target policy for optimization and a behaviour policy for exploration.

Within the RL/DRL approaches, we distinguish several key methods: value-based, policy-based, and actor-critic methods [28]. Value-based methods are generally more suitable for problems with discrete action spaces. For continuous action spaces, policy-based or actor-critic methods are typically preferred.

Value function learning leverages the value function to assess the quality of a given state when the agent follows a specific policy. Consequently, a robotic agent can refine its policy by exploring the action space to identify actions that maximize the estimated value function. Key Value methods are Q-learning, SARSA, Deep Q-Networks (DQN), Double Q-learning [29]. Q-learning, uses the maximum Q-value for the next state to update Q-values. It can lead to overestimation in some cases. Deep Q-Network (DQN) employs deep neural networks to estimate Q-values, enabling efficient decision-making in environments with complex and high-dimensional state spaces. It stabilizes training with experience replay and target networks. Double DQN mitigates overestimation bias by utilizing two separate Q-networks: one to determine the best action and another to evaluate its corresponding Q-value, leading to more accurate value estimations. SARSA is an on-policy method, meaning it updates the Q-values based on the actions that the agent actually takes while following its policy, as opposed to Q-learning, which is an off-policy method and updates the Q-values based on the maximum possible future reward.

Policy gradient directly optimizes the policy in a model-free manner by maximizing the accumulated reward. Policy-based algorithms, compared to value-based methods, typically provide better convergence and can learn stochastic policies. Examples of policy-based methods in Reinforcement Learning (RL) include Proximal Policy Optimization (PPO), Deterministic Policy Gradient (DPG), and Trust Region Policy Optimization (TRPO). They focus on directly optimizing the policy to improve decision-making, rather than learning a value function.

PPO is a policy-based method that improves the stability of policy updates using a clipped objective function. PPO is designed to handle large, complex environments and is widely used due to its efficiency and simplicity. TRPO is another policy-based method that ensures updates to the policy are within a *trust region* to prevent overly large changes that can destabilize learning. It uses a constrained optimization approach, making it more computationally expensive but stable and effective. DPG is a policy gradient method specifically designed for continuous action spaces. It learns a deterministic policy and utilizes a policy gradient approach to directly update the policy without needing a value function for actions.

Actor-Critic algorithms integrate two key components: an actor, responsible for determining actions, and a critic, which assesses the chosen actions' value. By optimizing both networks concurrently, these algorithms enhance stability and learning efficiency in reinforcement learning. Notable Actor-Critic approaches include Advantage Actor-Critic (A2C), Asynchronous Advantage Actor-Critic (A3C), Deep Deterministic Policy Gradient (DDPG), Twin Delayed Deep Deterministic Policy Gradient (TD3), and Soft Actor-Critic (SAC). The summary of Actor-critic algorithms is presented in Table 1.

The main challenges of applying reinforcement learning to manipulation tasks in robotics include the high dimensionality, the continuous action space, and the significant training time required [30]. Model-free approaches like PPO, DDPG and SAC are widely used in robot manipulation. These algorithms excel in learning continuous control policies for complex robotic arms and manipulators. In our research, we used CQL/SAC algorithm.

Table 1 Characteristics of Actor-critic algorithms

Algorithm	Type	Key Features	Action Space
A2C	Synchronous	Advantage function to reduce variance	Discrete and continuous
A3C	Asynchronous	Multiple parallel agents for stability and efficiency	Discrete and continuous
DDPG	Off-Policy	Deterministic policy, used for continuous action spaces	Continuous
TD3	Off_policy	Double Q-learning and target network smoothing to reduce bias	Continuous
SAC	Off-Policy	Includes entropy regularization for exploration	Continuous

3.1 CQL/SAC algorithm

SAC is a DRL algorithm built on the Actor-Critic framework and operates as an off-policy method. It is designed to overcome the stability and efficiency limitations present in earlier approaches. SAC utilizes the maximum entropy reinforcement learning paradigm, where the actor seeks to maximize both the expected reward and exploration by maximizing entropy [31].

By employing off-policy updates, SAC allows for faster learning and improved sample efficiency through experience replay, unlike on-policy methods like Proximal Policy Optimization (PPO), which require fresh data for each gradient update. In contrast, off-policy methods reuse past experiences, significantly enhancing learning efficiency. Unlike PPO and DDPG, SAC employs twin Q-networks alongside a separate Actor network, with entropy tuning mechanisms to improve both stability and convergence.

The integration of CQL promotes more stable learning by addressing the issue of Q-value overestimation, which is especially useful when dealing with offline or suboptimal data scenarios [13]. The advantage of combining Conservative Q-Learning (CQL) with Soft Actor-Critic (SAC) lies in the synergy between robust exploration and conservative value estimation. SAC excels at learning effective policies through its entropy-regularized framework, which promotes balanced exploration and exploitation. However, SAC can sometimes overestimate Q-values, especially in data-limited or offline scenarios. By introducing a conservative penalty, CQL mitigates this issue, resulting in improved training stability and policy robustness.

Compared to PPO and Twin Delayed Deep Deterministic Policy Gradient (TD3), the CQL/SAC model offers advantages in both training time and efficiency. While SAC requires more training time than PPO, it achieves more stable learning, and it converges faster than TD3 due to its entropy-based exploration [8], [32, 33]. Although CQL slightly increases training time, it significantly enhances learning stability [13]. Overall, CQL/SAC is more sample-efficient than PPO and better suited for real-world deployment than TD3, particularly in high-dimensional tasks and offline learning scenarios.

This CQL/SAC hybrid method leads to safer and more efficient decision-making, particularly in dynamic, continuous control tasks where real-world uncertainties and limited data are common challenges [25]. RL objective for SAC algorithm is:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim \rho_\pi}[(r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot\,|s_t))] \tag{1}$$

where $\mathcal{H}$ is the entropy:

$$\mathcal{H}\pi(\cdot\,|s_t) = \mathbb{E}[-\log(f_\pi(\cdot\,|s_t))] \tag{2}$$

with expectation operator $\mathbb{E}$, which denotes the averaging over all state-action pairs sampled from the trajectory distribution. A temperature parameter α is controlling the balance between exploration (higher entropy) and exploitation (higher reward). SAC can tune α automatically during training. Ultimately, this objective function aims to optimize the cumulative expected reward while simultaneously promoting exploration by maximizing policy entropy.

SAC follows an actor-critic architecture, where the actor explicitly models the policy, while the critic is solely responsible for guiding the actor's improvement. The critic's role is confined to the training phase, without directly influencing action selection during execution.

Two separate critic networks Q1 and Q2 are trained to minimize the Bellman error using the target Q value:

$$\hat{Q}_{\hat{\theta}_1,\hat{\theta}_2}(s_{t+1},a_{t+1}) = r_t + \gamma\mathbb{E}_{(s_{t+1}\sim D,a_{t+1}\sim\pi_\phi(.|s_{t+1}))}[\hat{Q}_{min} - \alpha\log(\pi_\phi((a_{t+1}|s_{t+1}))] \quad (3)$$

where the actor network $\pi_\theta(a|s)$ is trained to maximize the expected return while encouraging exploration using the entropy regularization term.

The target Q value incorporates the minimum of the two Q-values (to mitigate overestimation bias) and the entropy of the policy:

$$\hat{Q}_{min}(s_{t+1},a_{t+1}) = \min[\hat{Q}_{\hat{\theta}_1}(s_{t+1},a_{t+1}), \hat{Q}_{\hat{\theta}_2}(s_{t+1},a_{t+1})] \quad (4)$$

For each critic network, the Q-loss is computed as the mean squared error between the predicted Q-value and the target Q value:

$$J_Q(\theta_i) = \frac{1}{2}\mathbb{E}_{(s_t,a_t\sim D)}[\left(\hat{Q}_{\hat{\theta}_1,\hat{\theta}_2}(s_{t+1},a_{t+1}) - \hat{Q}_{\theta_1}(s_t,a_t)\right)^2] \quad (5)$$

The target networks are slowly updated copies of the critic networks. The target network does not have its own loss function; it serves to provide stable target values for training the critic network.

The policy-loss for actor-network is defined as:

$$J_\pi(\phi) = \mathbb{E}_{(s_t\sim D,a_t\sim\pi_\phi(.|s_t))}[\alpha\log(\pi_\phi(a_t|\, s_t)) - \min[(Q_{\theta_1}(s_t,a_t^\pi), Q_{\theta_2}(s_t,a_t^\pi)]] \quad (6)$$

A schematic view of SAC algorithm is presented in Fig. 1.

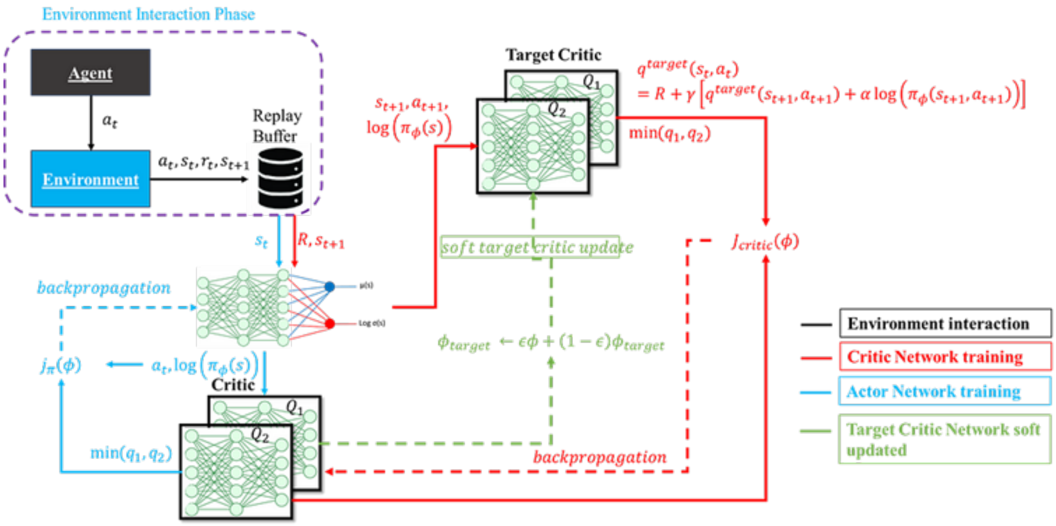


Fig. 1 A schematic view of the SAC algorithm [31]

In offline reinforcement learning (offline RL), the non-Lagrange version of the Conservative Q-Learning (CQL) method is utilized, as proposed in [13]. This approach is beneficial because it seamlessly integrates with existing continuous RL algorithms like Soft Actor-Critic (SAC) by introducing a regularization loss. By incorporating the CQL loss term into Eq. 5, the total Q-loss is expressed as:

$$J_Q^{total}(\theta_i) = J_Q(\theta_i) + \alpha_{cql}\mathbb{E}_{(s_t\sim D)}[log\sum_{a_t} exp(Q_{\theta_i}(s_t,a_t) - \mathbb{E}_{(a_t\sim D)}[Q_{\theta_i}(s_t,a_t)]) \quad (7)$$

where α_{cql} determines the degree to which the CQL loss is applied to the Q-loss. This penalizes actions that significantly deviate from the existing dataset, ensuring that the learned policy remains conservative in terms of exploration.

SAC algorithm is highly effective in continuous action spaces and complex tasks that require real-time adaptation, which is critical for industrial applications [9]. Due to these advantages, SAC can significantly contribute to the automation and optimization of pick-and-place operations in industry, making robotics more efficient and adaptable in dynamic production environments [8].

4. System description

The pick-and-place task requires a robotic arm to detect, grasp, and relocate an object to a target position efficiently. Key challenges include accurate perception using sensors (e.g., cameras, depth sensors), motion planning to compute smooth and collision-free trajectories for object grasping and placement, grasp stability to prevent object slippage, and dynamic adaptation to object variations.

This study focuses on enhancing the performance of RL agents in robotic pick-and-place tasks. To initiate this process, the robot interprets human gestures to trigger the intended pick-and-place actions. The block diagram of the system is presented in Fig. 2.

The cobot (myCobot320 by Elephant Robotics) is trained to perform pick-and-place tasks separately. The pick range is 10 cm × 15 cm, and the place range is identical but mirrored along the y-axis. The object's initial position is within the pick range, while the place position is randomly assigned within the place range. The trained policies are then combined to execute tasks sequentially in both simulation (PyBullet) and real-world scenarios. For real-world object detection, an Astra Pro 2 camera uses HSV color space, with tests conducted on green objects of 2 × 2.5 cm and 5 × 5 cm. Additionally, human collaboration is incorporated, allowing the robot to pause and resume operation based on hand state recognition using Google's MediaPipe library.

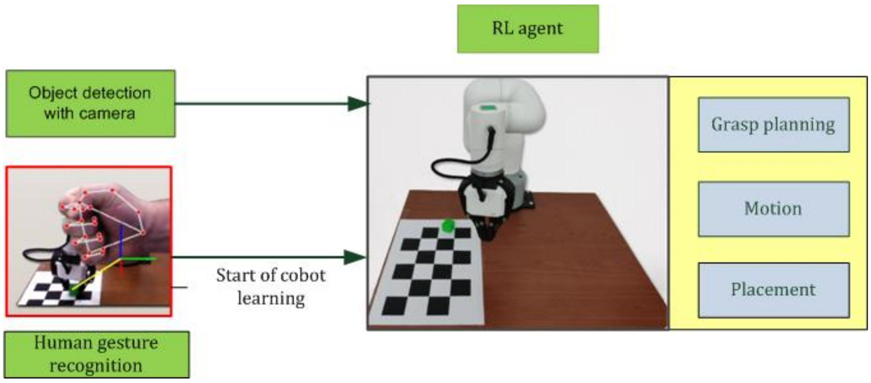


Fig. 2 A schematic view of Pick-and-place task in human-robot collaboration setting

4.1 Simulated and real-world platform

Robotic systems, with their high-dimensional and continuous action spaces, require extensive training for agents to develop optimal policies. However, online training can be expensive due to factors like the need for human oversight, potential risks of robot wear and damage, and limited access to training robots, making simulation-based training a practical alternative. Despite its benefits, simulation often suffers from low accuracy when applied to real-world tasks. Transferring policies from simulation to real-world robotics faces several challenges due to the sim-to-real gap, including [30, 34]:

- Model Inaccuracies: Simulations often oversimplify physics, neglecting factors like friction, sensor noise, and joint flexibility.
- Perception Discrepancies: Simulated sensors lack real-world noise and distortions, leading to poor generalization in tasks like vision-based navigation.
- Actuation Latency: Real actuators introduce delays and non-linearities not present in simulations.
- Data Distribution Shift: Policies trained in simulation may not generalize well to real-world variations.
- Safety and Robustness: Errors in real-world execution can damage hardware, and policies must handle unpredictable human interactions.
- Transfer Learning: Simulation-trained policies often need fine-tuning in the real world, which can be costly and inefficient.
- Lack of Real-World Data: Collecting real-world data for training RL policies is resource-intensive.

A major challenge in RL for robotics lies in bridging the Sim2Real gap, as transferring policies learned in simulators like Mujoco, PyBullet, or Gazebo, to real-world environments remains difficult. Addressing this challenge requires not only improving the robustness of RL algorithms but also developing more accurate and reliable simulators [35]. The development of standardized benchmarks and open-source environments, such as OpenAI Gym, Robosuite, and Isaac Gym, has accelerated research in RL for robotic manipulation. These advancements would provide suitable training environments for RL, enabling the creation of policies enabling them to perform effectively on physical robots.

This paper explores solutions to these challenges, focusing on leveraging RL to enable robots to autonomously learn and execute pick-and place tasks in both simulated (PyBullet environment), presented in Fig. 3 and real-world environments using myCobot320 by Elephant Robotics, presented in Fig. 4.

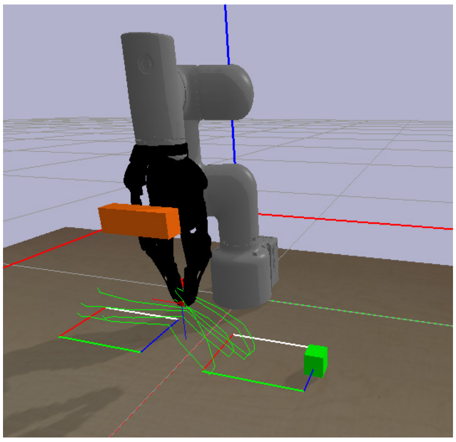


Fig. 3 Simulated environment (PyBullet)

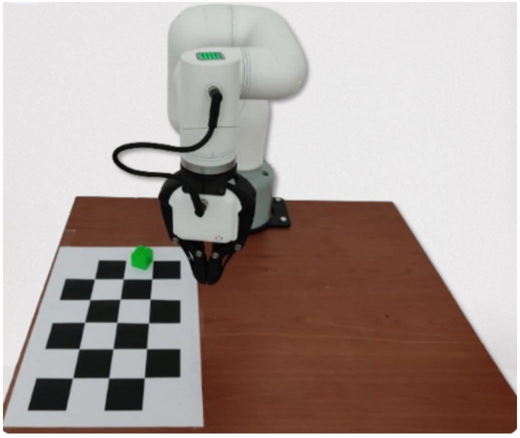


Fig. 4 Hardware settings for real environment

4.2 Human gesture recognition

Human collaboration is applied by stopping and resuming the work of real robot by recognizing the state of the hand by using Google's MediaPipe library. MediaPipe is an open-source framework developed by Google for building cross-platform, real-time perception pipelines, particularly in computer vision and machine learning applications. It utilizes graph-based processing, where modular components (calculators) efficiently handle tasks such as object detection, pose estimation, and gesture recognition.

Google's MediaPipe library can be utilized for gesture recognition by leveraging its pre-trained models for hand tracking. It detects key points on the hand, such as the position of each finger and the palm, allowing for precise gesture recognition in real-time. By integrating MediaPipe into robotic systems, gestures can be used as inputs to control the robot's actions, enabling intuitive human-robot interaction.

MediaPipe utilizes GPU acceleration and model optimization strategies to enable high-performance inference on edge devices, such as smartphones and embedded systems. Its architecture is designed for fast prototyping and seamless deployment of AI-powered applications across different platforms, while minimizing computational demands.

In the event of failure (due to unexpected human interventions beyond simple gesture-based commands), the real robot will perform unusual actions based on observations that are currently unseen. Typically, in such situations, we stop the execution.

4.3 RL agent

In our study, the RL agent leverages key components to learn effective object manipulation within a human-robot collaboration setting. By observing the state, taking actions, and receiving rewards, the agent refines its policy to enhance task performance, such as successful object picking and placement.

The state encapsulates essential details about the environment and the robot's condition, including the end-effector's position and orientation, the real object's position relative to the robot's gripper, the target placement position, the computed desired object position based on the end-effector and real object positions, and the gripper's state (open/closed), resulting in a state size 22 in total.

The agent operates with continuous actions, such as executing small translations and rotations of the end-effector, adjusting joint angles for articulated control, and managing the gripper's state (open/close) to grasp or release objects, resulting in a total size of 7 actions.

The reward function is designed to encourage efficient task completion while discouraging suboptimal behaviours. Positive rewards are given for actions such as successfully grasping the object, moving it closer to the target, and correctly placing it in the target area. Negative rewards (penalties) are applied for actions like collisions with obstacles, dropping the object before reaching the target, and unnecessary movements that delay task completion.

In the context of the Soft Actor-Critic (SAC) framework and Conservative Q-learning (CQL) on top of the framework, the policy network plays a crucial role in determining the actions the robot should take based on the current state [8]. We use two Multi-Layer Perceptron (MLP), one for policy (actor) and the other for Q-value (critic, Q1 and Q2). The structure of both networks is the same except for the input/output layer. The structure of MLP consists of 3 fully connected hidden layers with size of 256 and batch size of 32. The policy network takes the state (observations) as input of size 22 and gives action as output of size 7, representing the actions the robot should take. The Q-value network takes the states and actions as input of size 29 and gives Q-value of size 1.

Overall, this neural network architecture within the CQL/SAC algorithm enables the robot to learn a policy that maps states to actions in a continuous space, continuously optimizing its performance over time through reinforcement learning.

Hyperparameters include the state dimension for the policy (actor) set to 22 and the action dimension set to 7. The total number of training epochs is 300 for the place task and 500 for the pick task. Each epoch consists of 1000 steps, where a step represents a single interaction where the agent takes an action, receives a reward, and moves to the next state.

Other hyperparameters include a maximum of 40 steps per episode during evaluation and a training batch size of 32. Evaluation occurs after each epoch (set to 1), using 16 episodes to compute evaluation metrics. An episode starts from an initial state and ends upon reaching a goal, a time limit, or failure.

CQL and SAC-specific hyperparameters include automatic entropy tuning (set to true), CQL alpha (1), CQL temperature (1), discount factor (0.99), target smoothing coefficient tau (0.005), and entropy regularization alpha (0.2). The number of random samples for CQL loss estimation is 10, with CQL version 3. Learning rates are 1e-4 for the policy network and automatic entropy tuning, and 3e-4 for the Q-value network.

4.4 Evaluation metrics

Evaluation metrics in robot manipulation are essential for assessing the performance, efficiency, and reliability of robotic systems in handling various tasks. These metrics help quantify how well a robot interacts with objects, executes movements, and completes assigned tasks under different conditions. Some key evaluation metrics include Trial Success Rate (TSR), Task Completion Time, Grasp Success Rate, Path Efficiency, etc. [36].

In this study, we used TSR and average cumulative reward over multiple episodes, which represents the percentage of multi-step tasks completed with 100 % success [37].

5. Experimental design, results and discussion

To evaluate CQL/SAC's performance in high-dimensional continuous control tasks, we train a cobot on a pick-and-place problem requiring accurate object detection, grasping, transportation, and precise placement. Figs. 5 and 6 show end-effector positions and object positions for four trajectories and ten trajectories, where green color represents successful trajectory/object placement and red color is for failed trajectory/object placement.

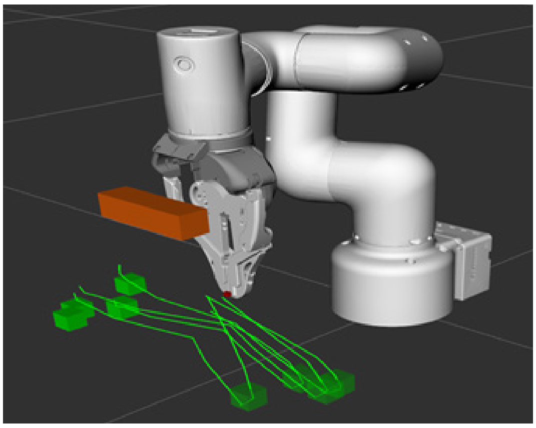


Fig. 5 End-effector positions for four trajectories

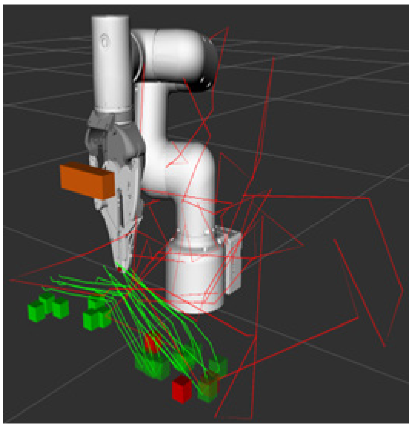


Fig. 6 End-effector positions for ten trajectories

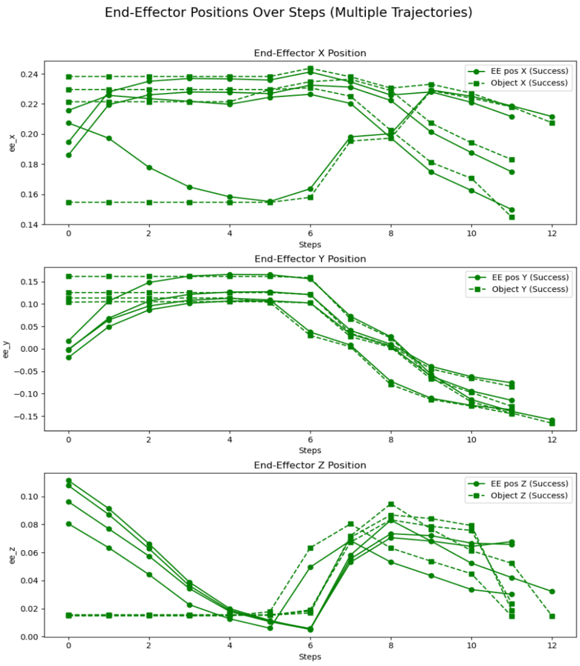


Fig. 7 Four successful trajectories, End-effector/object positions

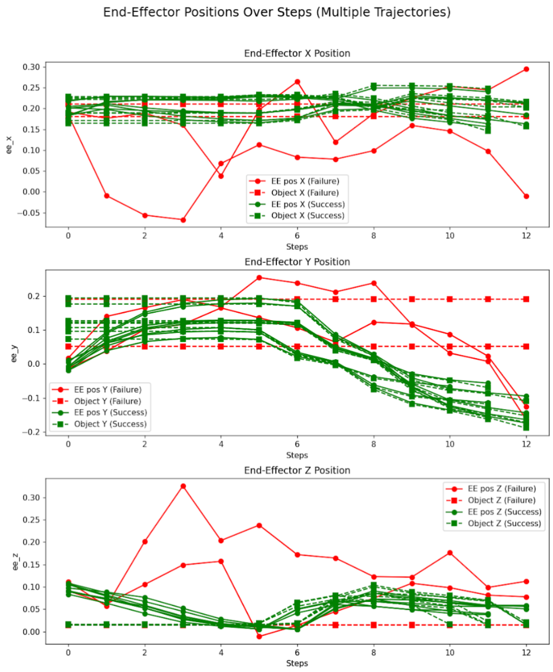


Fig. 8 Ten mixed (successful and failed) trajectories, End-effector/object positions

Fig. 7 and Fig. 8 presents trajectories of end-effector (dots) positions and object coordinates (squares) as functions of steps for 4 and 10 trajectories, respectively.

Fig. 9 and Fig. 10 illustrate the best performance metrics of picking and the placing tasks for evaluation and training. Neural networks with such trained weights are used for the real robot for pick and place tasks respectively. Pick task is trained for 500k steps and place task for 300k steps. From Fig. 9 can be seen that pick task gets 100 % success rate after 280k episodes, and mean rewards during training and evaluation remain stable. The similar results can be concluded from Fig. 10 for place task where after around 100k steps (1 epoch) we get the best performance regarding evaluation success rate and trained/evaluated reward remains similar.

Figs. 11 and 12 presents comparisons between different runs of pick and place tasks respectively. From Fig. 11 we started with 100k steps (run called `pick_0`) for training for pick task. We can see that it didn't get a successful eval success rate (obtained success rate is 0), even though other parameters were looking promising. The same happened for other runs (`pick_{1,2,3}`) respectively, where we can also see that RL is not deterministic. After expanding the training time from 100k to 500k we obtained the `pick_best` run.

Fig. 9 The best performances during training and evaluation for pick task

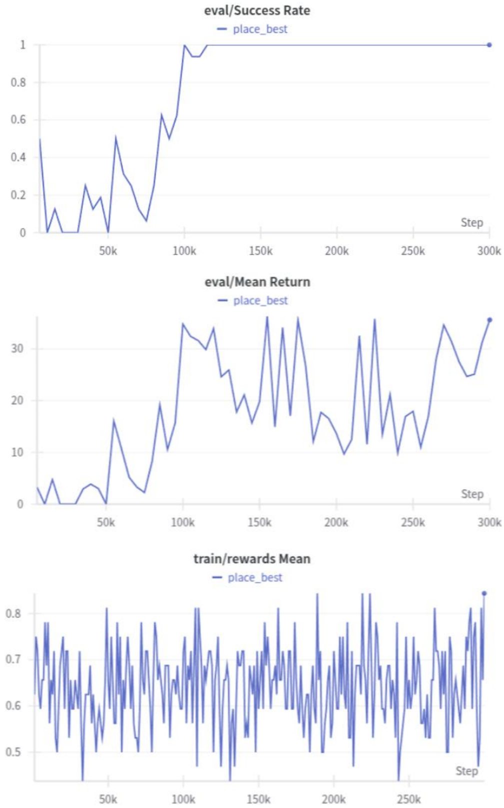


Fig. 10 The best performances during training and evaluation for place task

Fig. 11 The performance comparison between different runs for pick task

Fig. 12 The performance comparison between different runs for place task

Similarly, for place task, we varied the total number of steps for different runs, from 130k (`place_0` run), 250k steps (`place_1` run) and 300k `place_best` run, which can be seen on Fig. 11. We observed that although the highest success rate was achieved during evaluation, stability throughout the evaluation steps was not maintained. Therefore, we extended the training time to 300k episodes.

The success rate achieved was 100% in simulation and 80% on the real robot, evaluated over 100 episodes.

6. Conclusion

In this research, we demonstrated the effectiveness of combining CQL/SAC for robotic manipulation in human-robot collaboration settings. By leveraging the combination of CQL and SAC, the robot successfully learned optimal policies for executing intricate tasks, such as object picking and placement. CQL ensures more conservative value estimation, preventing suboptimal actions that could arise due to overestimation, while SAC facilitates efficient exploration and adaptation in uncertain environments. The results show that this hybrid approach significantly enhances the performance and reliability of robotic manipulation, enabling robots to effectively collaborate with humans in shared environments. We assessed this approach through cobot manipulation experiments, demonstrating the transferability of the learned policy from simulation to real-world settings without additional training. This was validated through real robot experiments, confirming that the integration of CQL with SAC enables safer and more reliable policy deployment in human-robot collaborative scenarios.

Currently, the framework is designed for pick-and-place tasks, but future work will focus on expanding its capabilities to a wider range of robotic manipulation tasks. Additionally, future research should explore multi-object detection, enabling the system to handle various shapes in a more dynamic industrial setting. Further advancements will also integrate diverse interaction modalities, including voice commands and text-based collaboration powered by Generative AI, such as Large Language Models, to enhance human-robot communication and adaptability.

Acknowledgement

This research is supported in scope of projects "Research and Development of Collaborative Intelligence in Service Robots for Industrial Applications", funded by Federal Ministry of Education and Science, Bosnia and Herzegovina and "Smart factory enabled by artificial intelligence at the edges of a distributed network - a prerequisite for Industry 5.0", funded by Ministry of education and science of TK, Bosnia and Herzegovina.

References

[1] Javernik, A., Buchmeister, B., Ojsteršek, R. (2022). Impact of Cobot parameters on the worker productivity: Optimization challenge, *Advances in Production Engineering & Management,* Vol. 17, No 4, 494-504, doi: 10.14743/apem2022.4.451.

[2] Banjanovic-Mehmedovic, L., Karabegovic, I., Jahic, J., Omercic, M. (2021). Optimal path planning of a disinfection mobile robot against COVID-19 in a ROS-based research platform, *Advances in Production Engineering & Management*, Vol. 16, No. 4, 405-417, doi: 10.14743/apem2021.4.409.

[3] Li, Y.C., Wang, X. (2024). Human-robot collaboration assembly line balancing considering cross-station tasks and the carbon emissions, *Advances in Production Engineering & Management,* Vol. 19, No 1, 31-45, doi: 10.14743/apem2024.1.491.

[4] Liu, Q., Liu, Z., Xiong, B., Xu, W., Liu, Y. (2021). Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function, *Advanced Engineering Informatics,* Vol. 49, Article No. 101360, doi: 10.1016/j.aei.2021.101360.

[5] Peters, J., Lee, D.D., Kober, J., Nguyen-Tuong, D., Bagnell, J.A., Schaal, S. (2016). Robot learning, In: Siciliano, B., Khatib, O. (eds.), *Springer handbook of robotics,* Springer, Cham, Switzerland, 357-398, doi: 10.1007/978-3-319-32552-1_15.

[6] Xing, H.R. (2024). Optimizing human-machine systems in automated environments, *International Journal of Simulation Modelling*, Vol. 23, No. 4, 716-727, doi: 10.2507/IJSIMM23-4-CO19.

[7] Ojstersek, R., Javernik, A., Buchmeister, B. (2021). The impact of the collaborative workplace on the production system capacity: Simulation modelling vs. real-world application approach, *Advances in Production Engineering & Management*, Vol. 16, No. 4, 431-442, doi: 10.14743/apem2021.4.411.

[8] Haarnoja, T., Zhou, A., Abbeel, P., Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, *ArXiv,* doi: 10.48550/arXiv.1801.01290.

[9] Gu, S., Holly, E., Lillicrap, T., Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates, In: *Proceedings of 2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 3389-3396, doi: 10.1109/ICRA.2017.7989385.
[10] Karabegović, I., Banjanović-Mehmedović, L. (2021). *Service robots: Advances and applications*, Nova Science Publisher, New York, USA.
[11] Kroemer, O., Niekum, S., Konidaris, G. (2021). A review of robot learning for manipulation: Challenges, representations, and algorithms, *ArXiv,* doi: 10.48550/arXiv.1907.03146.
[12] Mahler, J., Matl, M. Satish, V., Danielczuk, M., DeRose, B., McKinley, S., Goldberg, K. (2019). Learning ambidextrous robot grasping policies, *Science Robotics,* Vol. 4, No. 26, Article Id. eaau4984, doi: 10.1126/scirobotics.aau4984.
[13] Kumar, A., Zhou, A., Tucker, G., Levine, S. (2020). Conservative Q-learning for offline reinforcement learning, *ArXiv,* doi: 10.48550/arXiv.2006.04779.
[14] Liu, Y., Wang, C., Zhao, C., Wu, H., Wei, Y. (2024). A soft actor-critic deep reinforcement-learning-based robot navigation method using LiDAR, *Remote Sensing*, Vol. 16, No. 12, Article No. 2072, doi: 10.3390/rs16122072.
[15] Kosaraju, D. (2024). Hierarchical reinforcement learning: Structuring decision-making for complex AI tasks, *International Journal of Science and Healthcare Research,* Vol. 7, No. 2, 485-491, doi: 10.52403/ijshr.20220468.
[16] Levine, S., Finn, C., Darrell, T., Abbeel, P. (2016). End-to-end training of deep visuomotor policies, *ArXiv,* doi: 10.48550/arXiv.1504.00702.
[17] Mohammed, M.Q., Chung, K.L., Chyi, C.S. (2020). Pick and place objects in a cluttered scene using deep reinforcement learning, *International Journal of Mechanical and Mechatronics Engineering,* Vol. 20, Vol. 4, 50–57.
[18] Han, D., Mulyana, B., Stankovic, V., Cheng, S. (2023). A survey on deep reinforcement learning algorithms for robotic manipulation, *Sensors*, Vol. 23. No. 7. Article No. 3762, doi: 10.3390/s23073762.
[19] Liu, D., Wang, Z., Lu, B., Cong, M., Yu, H., Zou, Q. (2020). A reinforcement learning-based framework for robot manipulation skill acquisition, *IEEE Access,* Vol. 8, 108429-108437, doi: 10.1109/ACCESS.2020.3001130.
[20] Al-Selwi, H.F., Aziz, A.A., Abas, F.S., Zyada, Z. (2021). Reinforcement learning for robotic applications with vision feedback, In: *Proceedings of the 2021 IEEE 17th International Colloquium on Signal Processing & Its Applications (CSPA)*, Langkawi, Malaysia, 81-85, doi: 10.1109/CSPA52141.2021.9377292.
[21] Yamada, J., Lee, Y., Salhotra, G., Pertsch, K., Pflueger, M., Sukhatme, G.S., Lim, J.J., Englert, P. (2020). Motion planner augmented reinforcement learning for robot manipulation in obstructed environments, *ArXiv*, doi: 10.48550/arXiv.2010.11940.
[22] Zhan, A., Zhao, P., Pinto, L., Abbeel, P., Laskin, M. (2020). A framework for efficient robotic manipulation, *ArXiv.*
[23] Kwon, G., Kim, B., Kwon, N.K. (2024). Reinforcement learning with task decomposition and task-specific reward system for automation of high-level tasks, *Biomimetics*, Vol. 9, No. 4, Article No. 196, doi: 10.3390/biomimetics9040196.
[24] Zhao, T., Wang, M., Zhao, Q., Zheng, X., Gao, H. (2023). A path-planning method based on improved soft actor-critic algorithm for mobile robots, *Biomimetics,* Vol. 8, No. 6, Article No. 481, doi: 10.3390/biomimetics8060481.
[25] Yadav, S.P., Nagar, R., Shah, S.V. (2024). Learning vision-based robotic manipulation tasks sequentially in offline reinforcement learning settings, *Robotica,* Vol. 42, No. 4, 1715-1730, doi:10.1017/S0263574724000389.
[26] Liu, A.Y., Yue, D.Z., Chen, J.L., Chen, H. (2024). Deep learning for intelligent production scheduling optimization, *International Journal of Simulation Modelling*, Vol. 23, No. 1, 172-183, doi: 10.2507/IJSIMM23-1-CO4.
[27] Wei, Z.H., Yan, L., Yan, X. (2024). Optimizing production with deep reinforcement learning, *International Journal of Simulation Modelling*, Vol. 23, No. 4, 692-703, doi: 10.2507/IJSIMM23-4-CO17.
[28] Wang, Y., Friderikos, V. (2020). A survey of deep learning for data caching in edge network, *Informatics,* Vol. 7. No. 4, Article No. 43, doi: 10.3390/informatics7040043.
[29] Liu, Z., Liu, Q., Xu, W., Wang, L., Zhou, Z. (2022). Robot learning towards smart robotic manufacturing: A review, *Robotics and Computer-Integrated Manufacturing,* Vol. 77, Article No. 102360, doi: 10.1016/j.rcim.2022.102360.
[30] Lobbezoo, A., Qian, Y., Kwon, H.-J. (2021). Reinforcement learning for pick and place operations in robotics: A survey, *Robotics,* Vol. 10, No. 3, Artice No. 105, doi: 10.3390/robotics10030105.
[31] Sharifi, N. Mathematical foundation underpinning reinforcement learning, from *https://ai.gopubby.com/mathematical-foundation-underpinning-reinforcement-learning-34b304890c33*, accessed January 3, 2025.
[32] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). Proximal policy optimization algorithms, *ArXiv,* doi: 10.48550/arXiv.1707.06347.
[33] Fujimoto, S., Hoof, H., Meger, D. (2018). Addressing function approximation error in actor-critic methods, *ArXiv,* doi: 10.48550/arXiv.1802.09477.
[34] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world, In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* Vancouver, Canada,, 23-30, doi: 10.1109/IROS.2017.8202133.
[35] Zhu, W., Guo, X., Owaki, D., Kutsuzawa, K., Hayashibe, M. (2023). A survey of sim-to-real transfer techniques applied to reinforcement learning for bioinspired robots, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 34, No. 7, 3444-3459, doi: 10.1109/TNNLS.2021.3112718.
[36] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,*The International Journal of Robotics Research,* Vol. 37, No. 4-5, 421-436, doi: 10.1177/0278364917710318.
[37] Hundt, A., Killeen, B., Greene, N., Wu, H., Kwon, H., Paxton, C., Hager, G.D. (2020). "Good Robot!": Efficient reinforcement learning for multi-step visual tasks with sim-to-real transfer, *IEEE Robotics and Automation Letters*, Vol. 5, No. 4, 6724-6731, doi: 10.1109/LRA.2020.3015448.

APEM journal

Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 18–28
https://doi.org/10.14743/apem2025.1.524

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Mass customization in practice: Strategic implementation and insights from Polish small and medium sized enterprises

Patalas-Maliszewska, J.[a,*], Kowalczewska, K.[b], Pajak, G.[a]

[a]Institute of Mechanical Engineering, University of Zielona Góra, Zielona Góra, Poland
[b]Doctoral School of Exact and Technical Sciences, University of Zielona Góra, Zielona Góra, Poland

A B S T R A C T

Implementing a mass customization (MC) strategy in manufacturing enterprises presents an ongoing challenge for both managers and researchers. To remain competitive, managers must consider adopting advanced technologies associated with Industry 4.0 and 5.0 (I4.0/5.0). This study seeks to identify solutions that support strategic decision-making aimed at enhancing the level of MC implementation. The paper begins with a literature review focused on the adoption of MC strategies within European manufacturing enterprises. It then presents findings from a questionnaire-based survey conducted among more than 100 small and medium-sized enterprises (SMEs) in Poland's automotive sector. Statistical analysis, including correlation coefficients, was used to evaluate the data. The results indicate that consumer participation in the product design process is the key driver of successful MC strategy implementation in the surveyed SMEs. Furthermore, managers recognized strong correlations between the adoption of I4.0/5.0 technologies—such as automated machinery and real-time data usage—and higher levels of MC capability. The benefits of implementing MC strategies, including increased production flexibility and waste reduction, were also highlighted. The findings offer general insights applicable to SMEs in the automotive industry.

A R T I C L E I N F O

Keywords:
Mass castomisation strategy;
Small and medium sized manufacturing enterprises;
Consumer participation;
Production flexibility;
Industry 4.0/5.0 technologies;
Automotive industry

**Corresponding author:*
J.Patalas-Maliszewska@iim.uz.zgora.pl
(Patalas-Maliszewska, J.)

Article history:
Received 18 December 2024
Revised 14 April 2025
Accepted 19 April 2025

1. Introduction

Designing and implementing an appropriate production strategy requires managers to have adequate knowledge about the potential effects and possible scenarios of enterprise development. Mass customization (MC) involves the production of various product variants while reducing the costs of tools and equipment, minimizing changes in production processes, machinery, the number of employees, and improving production flexibility and quality [1]. Currently, the strategy of MC and personalization is one of the challenges related to the need to implement the Industry 4.0/Industry 5.0 (I4.0/5.0) concept in enterprises. MC can be treated as a production strategy that combines push and pull production paradigms to achieve a core competence [2]. Customization can be treated as the process of adapting a product or service to the customer's own needs, most often with support from information systems such as product configurators. Customized products currently respond to customer expectations and enable companies to gain a competitive advantage. An open research question remains: How can the level of MC be defined and measured, and what indicators signal the success of implementing this production strategy? According to [2], MC research has been unexplored, and therefore manufacturing companies have low competence in implementing MC. Customers increasingly demand personalized, high-quality products at competitive prices [3-4]. To identify the research gap, a preliminary literature review (Table 1) was conducted with the aim of defining key factors influencing the degree of mass customization (MC) strategy implementation, based on empirical evidence from European manufacturing enterprises.

Table 1 A literature review of current factors determining a level of MC strategy implementation in European Enterprises

MC realizing in the manufacturing enterprises	European country	Factors	Source
Small and medium enterprises, Mechanical products	Italy	Product platform development, IT-based product configuration, Group technology	[5]
Clothing (Burberry), Furniture (IKEA) and toy (LEGO) industries	UK, Denmark, Sweden	Appropriate business and marketing strategy, Operational management in accordance with sustainable production and technological development	[6]
Machine building industry (FENDT)	Germany	Highly flexible assembly line	[7]
Automotive and agricultural industries (Fiat powertrain technologies)	Italy	Use real-time data	[8]
Automotive industry (assembly a diesel particulate filter (DPF))	-	Use augmented reality technology, Dynamic production rescheduling	[9]
Semiconductor industry	-	Robot utilization, Simulation and statistical analyses	[10]
Electronic industry (laptops/PCs)	-	Product configuration mechanism	[11]
Assembly industry	-	Modeling and simulation techniques of production processes	[12]
Industrial enterprises	Czech Republic	Industry 4.0 (Flexible processes, Artificial intelligence-based solutions, Automation, Robotics, e-commerce, 3D printing, Flexible manufacturing)	[13]
Manufacturing sector	-	Collaboration networks, Business agility, Digital supply chain, Use of I4.0 technologies	[14]
Electronic industry, Automotive industry	EU	High level of modular design, Remanufacturing of modular products	[15]
Automotive industry	-	Clear description of configuration options provided to customers, Flexible manufacturing system and processes	[16]

Table 1 summarizes the findings of the literature review concerning the implementation of mass customization (MC) strategies in European manufacturing enterprises. Specifically, it presents (1) the types of European industries analyzed, (2) the countries in which these industries are located, and (3) the key factors identified as influencing the level of MC strategy implementation. To the best of our knowledge, and based on the current state of the literature (as presented in Table 1), no previous studies have explicitly investigated how these factors contribute to increasing the level of MC implementation within European manufacturing enterprises in the automotive sector. Furthermore, there is a lack of research addressing the impact of emerging technologies—specifically those associated with the Industry 4.0/5.0 paradigm—on the degree of MC implementation. While prior studies (e.g., [2]) suggest that MC is inherently linked to technological advancement, the reviewed literature (Table 1) does not provide empirical evidence on the relationship between the adoption of Industry 4.0/5.0 technologies and the enhancement of MC capabilities in manufacturing firms.

Therefore, in this paper we examine the following research questions: (1) What key factors influence increasing the level of MC in European Manufacturing Enterprises? (2) What kind of

I4.0/5.0 technologies impact the MC level, and (3) How can these relationships influence future research trends in MC research?

The I4.0/5.0 technologies in the context of MC strategy realization can be distinguished as robotics, simulation and integration, cyber-physical systems (CPSs), Internet of things (IoT), cloud computing, automated machines, AI, augmented reality (AR) and virtual reality (VR), cybersecurity, human-machine interaction (HMI), and finally human-robot collaboration (HRC). The usage of robots generally enables flexible and efficient production. But in the case of MC, it requires operators to perform responsible tasks to adapt production to individual customer requirements. Simulation of production processes allows presenting different scenarios to improve decision-making in an MC strategy [17]. CPSs integrate technologies with physical systems in order to increase the automation of MC production [18]. IoT-based solutions facilitate maintaining the operating state of the system within MC production [19]. Cloud computing allows manufacturers to respond to the fact that more and more customers are willing to participate in the design process, and it enables MC production to be flexible and scalable [20]. Automated machines enable production tailored to the customer's needs. AI-based tools provide visualization of processes, their monitoring and control, configuration of products, quality assurance, and real-time data classification to maximize process efficiency [21]. AR and VR for MC production enable product visualization for customers [9]. Cybersecurity is a set of necessary technologies in MC production for the secure collection of customer data [22]. HMI in MC enables more effective delivery of products to market [23].

This paper analyses and models the implementation of the mass customization (MC) strategy in European manufacturing, using Polish small and medium-sized manufacturing enterprises (SMEs) as a representative case study. Furthermore, it aims to identify and define the relationships between the adoption of Industry 4.0/5.0 technologies and the achieved level of MC implementation. The originality of this study is as follows:

- This study establishes the relationships between the I4.0/5.0 technologies implementation and the increased level of MC strategy for European SMEs in the automotive industry.
- It provides practical insights into MC through empirical research in over 100 European manufacturing SMEs from Poland.
- This study determines the benefits of implementing the MC strategy in the automotive industry.
- Recommendations for managers in the automotive manufacturing industry were formulated regarding necessary actions to increase the level of MC strategy.

2. Materials and methods

2.1 Questionnaire surveys on MC

The usage of a tool such as a questionnaire enables the collection of research data among many manufacturing companies in a short time, as the respondents provide answers to specific questions in writing, either traditionally or electronically. The questionnaire used in this study concerned the implementation of the MC strategy in manufacturing enterprises, factors determining the level of MC strategy realization, and the influence of the implementation of I4.0 technologies on the level of the MC strategy.

Firstly, the level of the MC strategy was defined. In our research, the MC strategy implementation level is classified according to the Technology Readiness Levels (TRL) classification. TRL is a classification that allows the determination of the technological maturity of a product, process, or service—from the creation of an idea and basic research, through conceptual and laboratory work corresponding to industrial research, to creating a prototype as part of development work, and finally, to a ready-made solution applicable in practice. TRL determines what validation activities have already been performed and what still need to be done [24]. According to TRL, five levels of implementation of the MC strategy were defined, depending on the answers given in the conducted surveys. Level I (TRL levels 1-2) is achieved if the respondent declares knowledge of the MC definition.

Level II (TRL levels 3 and 4) is achieved if, in a company, customized orders are tailored to the individual needs of customers. Level III (TRL level 5) concerns analytical and experimental confirmation of critical functions or conscepts of the technology. Level IV (TRL levels 6 and 7) refers to the stage when the technology components or its basic subsystems have been integrated, and the company plans to implement the MC strategy. Level V (TRL levels 8 and 9) is achieved when the company defines the methods for realizing the MC strategy.

Next, based on the literature research results (Table 1), it was possible to define the factors for European Manufacturing Enterprises that influence the level of the MC strategy (Table 2) in our questionnaire.

Table 2 Factors describing MC strategy

Factors describing MC strategy	Abbreviations
Consumer participation in the design process	C1
Customer participation in the product design	C2
Integration of customer preferences	C3
Product configuration mechanisms	C4
The ability to handle multiple process variations at different stages of production	C5
Highly flexible assembly line	C6
Using modeling and simulation techniques of production processes	C7
The ability to use real-time data to make efficient, quick decisions to assembly line	C8

Table 3 Industry 4.0/5.0 Technologies – Factors of implementation

Industry 4.0 technology	Abbreviations
Robotics	I1
Product simulations	I2
Material simulations	I3
Production processes simulations	I4
Cyber-physical systems (CPSs)	I5
Systems integration	I6
Industrial Internet of Things (IIoT)	I7
Cloud computing	I8
Automated machines	I9
Artificial Intelligence tools	I10
AR and VR	I11
HMI and HRC	I12
Cybersecurity	I13

Subsequently, to address the second research question, the survey incorporates items related to the implementation of selected I4.0/5.0 technologies, as outlined in Table 3.

To investigate how the identified relationships may shape future research directions in the field of mass customization (MC), empirical research was conducted among the surveyed enterprises to evaluate the benefits associated with the implementation of the MC strategy within the automotive industry (Table 4).

The questionnaire is structured to include an introductory section outlining the objectives and scope of the survey, followed by four distinct modules.

Table 4 Benefits of MC strategy implementation

Factors	Abbreviations
Annual sales increase up to 10 %	E1
Annual sales increase from 11 to 25 %	E2
Annual sales increase above 25 %	E3
Increasing the flexibility of the work surface	E4
Increasing production flexibility	E5
Reducing the amount of waste	E6
Reducing electricity consumption	E7
More effective analysis of large data sets	E8
Decentralization of decision-making	E9

The first module comprises a set of open- and closed-ended questions designed to collect essential company-related information from the respondent. This includes the company's registered office and operational location, the number of employees (enabling enterprise size classification), the primary product portfolio, and the respondent's position within the organization. These elements allow for an assessment of the respondent's familiarity with the company's management strategy and development policy.

The second module is completed by respondents whose companies have implemented a mass customization (MC) strategy. It covers questions related to expenditures incurred during implementation, methods of gathering customer preferences and requirements, approaches and outcomes of MC implementation, the level of satisfaction associated with MC deployment, the I4.0/5.0 technologies utilized to support MC, and future plans concerning the extension of MC applications within the enterprise.

The third module targets respondents from companies that have not adopted the MC strategy. It includes a multiple-choice question addressing the reasons for non-implementation, alongside the possibility to provide an open-ended explanation and indicate whether implementation is planned for the future.

The fourth module explores the perceived significance of I4.0/5.0 technologies in achieving the intended outcomes of MC strategy implementation.

2.2 Research group and data collection

The empirical study was conducted using a structured questionnaire composed of closed, multiple-choice questions. Data collection took place between January 10 and August 31, 2023, through both face-to-face interviews (29 %) and telephone surveys (71 %).

The survey followed a sample-based research design and gathered responses from 153 European manufacturing enterprises operating in Poland within the automotive sector. The sample consisted of 117 small and medium-sized enterprises (SMEs; defined as employing up to 249 individuals) and 36 large enterprises. The selection of the automotive industry as the focus of analysis is justified by its strategic importance to the European economy, accounting for nearly 7 % of the region's gross domestic product (GDP). Furthermore, the automotive sector employs approximately 13.8 million individuals in the European Union, which represents 6.1 % of the total workforce.

Notably, collaboration among partners within the automotive industry remains an open research question [25], further validating the relevance of this sector for investigation.

This study places particular emphasis on SMEs, as they constituted over 76 % of the total research sample.

The research sample may be considered representative. According to data from the Polish Central Statistical Office, 3,954 enterprises were registered in 2022 as operating in the automotive industry. The obtained sample size of 153 exceeds the minimum required number of 145 enterprises, calculated at a 95 % confidence level, with a proportion (p) of 0.5 and a maximum margin of error of 8 %. The required minimum sample size was determined using the standard formula (Eq. 1):

$$N_{min} = \frac{N_p(\alpha^2 \cdot f(1-f))}{N_p \cdot e^2 + \alpha^2 \cdot f(1-f)} \tag{1}$$

where N_{min} – minimum sample size, N_p – population size, α – confidence interval, f – fraction size, e – assumed maximum error.

Moreover, the research was conducted across the entire territory of Poland, ensuring representation proportional to the number of companies registered in each voivodeship.

2.3 Research model

Based on the results of in-depth interviews with 117 SMEs from the automotive industry, a research model (Fig. 1) was developed and analysed using the correlation and regression method to estimate the level of Mass Customization (MC) strategy implementation in manufacturing enterprises. The survey instrument used for testing the model was constructed by defining appropriate measurement scales to assess both the impact of MC strategy implementation and the influence of Industry 4.0/5.0 technologies on its realization.

The factors describing the level of MC strategy realization in Polish automotive SMEs were derived from structured feedback and are listed in Table 2. Respondents assessed the importance of each factor (c1-c8) using a binary scale: factor0 – *not very important*, factor1 – *very important* for increasing the level of MC implementation.

Similarly, the perceived benefits of MC strategy implementation (Table 4) were evaluated based on company experiences in 2022, with respondents indicating whether each benefit (E1-E9) was considered *not very important* (factor0) or *very important* (factor1).

The overall level of MC strategy implementation in the surveyed companies was classified into five levels according to the Technology Readiness Level (TRL) framework.

Subsequently, our research also examined the implementation of Industry 4.0/5.0 technologies (Table 3). The assessment of these technologies was based on structured survey responses. Respondents were asked to indicate whether their company applies specific I4.0/5.0 technologies (I1-I12), and to evaluate the perceived significance of each technology in the context of MC strategy implementation. The following binary scale was applied: factor0 – *not very important*, factor1 – *very important*.

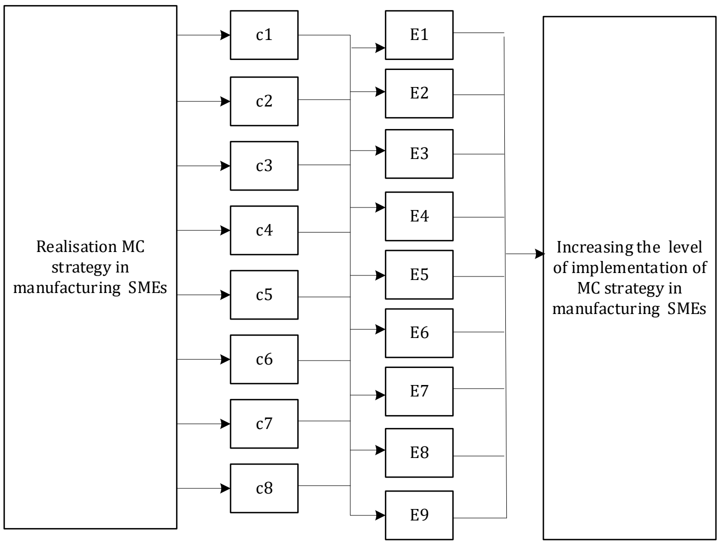


Fig. 1 Research model

3. Results

3.1 Results from the surveyed enterprises

When addressing the first research question, it can be confirmed that the factors identified in the relevant literature (Table 1) were also recognized by the surveyed enterprises. Among Polish manufacturing SMEs, 55.56 % of companies enable customer participation in the product design process, while 52.99 % of enterprises report both customer participation and integration of customer preferences. Product configuration mechanisms and the ability to utilize real-time data for efficient and rapid decision-making on the assembly line are implemented by 44.44 % of surveyed companies. A highly flexible assembly line is employed by 41.88 % of enterprises, simulation techniques for production processes are utilized by 30.77 %, and the ability to manage multiple process variations across different production stages is reported by 18.80 % of SMEs.

Furthermore, the empirical research confirms that I4.0/5.0 technologies (as listed in Table 3) are actively implemented among the surveyed manufacturing SMEs. Automated machines are utilized by 41.88 % of the surveyed enterprises, while systems integration is present in 35.04 % of enterprises. Product simulations, production process simulations, and cybersecurity solutions are adopted by 29.06 % of respondents. Material simulations are implemented by 23.93 % of enterprises. HMI and HRC are utilized by 22.22 % of SMEs, robotics by 21.37 %, and CPSs by 17.09 %, respectively. However, cloud computing is employed by only 7.69 % of SMEs, IIoT by 5.98 %, artificial intelligence tools by 3.42 %, and AR and VR technologies by 1.71 %, respectively.

These findings indicate that the most commonly adopted I4.0/5.0 technologies among European SMEs in Poland include automated machines, systems integration, simulations of products and production processes, as well as cybersecurity. Therefore, to address the first and second research questions, statistical analysis was conducted using correlation coefficients to identify the strength of relationships between the identified factors and the level of MC strategy implementation.

3.2 Analysis MC strategy realizing in Polish manufacturing SMEs

The correlation analysis is a statistical method used to examine the strength and direction of a linear relationship between two variables, quantified by the correlation coefficient r, which takes values in the interval $\langle -1, 1\rangle$. A value of −1 indicates a perfectly negative linear relationship, while +1 denotes a perfectly positive linear relationship. A value of 0 signifies the absence of a linear correlation (Bobko, 2001). The Pearson correlation coefficient is calculated according to the formula Eq. 2:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{2}$$

where x_i and y_i are the values of the variables x and y, respectively, and $\bar{x}$, $\bar{y}$ are the mean values of these variables.

During the statistical analysis, the correlation between the factors related to the realization of the MC strategy and the increase in the level of MC strategy implementation was examined. The analysis was conducted using Statistica version 13.3 (StatSoft Polska Sp. z o.o., Kraków, Poland).

The results of the correlation analysis are presented in Table 6. The table includes the following indicators: r^2 – coefficient of determination, t – value of the t-statistic testing the significance of the correlation coefficient, and p – probability value (significance level).

A very strong positive correlation was observed between the increase in the level of MC strategy and consumer participation in the design process (r = 0.8278). Additionally, significant correlations were identified between the increase in MC strategy and both customer involvement in product design (r = 0.7861) and the integration of customer preferences (r = 0.7861). In contrast, a weak correlation was found between the increase in MC strategy and the automation of production process planning as well as the use of simulation techniques.

These findings clearly indicate that the key factors influencing the enhancement of MC strategy in Polish manufacturing SMEs include consumer participation in the design process and product design, the ability to understand customer preferences, the utilization of product configuration mechanisms, production process control, and the application of real-time data.

In response to the second research question concerning the impact of advanced technologies on the advancement of MC strategy implementation, the relationships between Industry 4.0 technologies adopted and the increase in MC strategy level were analyzed and are presented in Table 6.

A strong correlation was observed between the increase in the level of MC strategy and the use of automated machines, such as 3D printers or autonomous processing stations (r = 0.6285), as well as with systems integration (r = 0.5438). Consequently, one of the key challenges in advancing the MC strategy is undoubtedly the automation of unique and customized processes.

Table 5 Correlations between the factors describing the realization of the MC strategy and the increase in the level of MC strategy in the automotive industry

Relations	Correlation	r^2	t	p
c1/MC	0.8278	0.6853	15.8271	0.0000
c2/MC	0.7861	0.6180	13.6419	0.0000
c3/MC	0.7861	0.6180	13.6419	0.0000
c4/MC	0.6622	0.4386	9.4792	0.0000
c5/MC	0.3563	0.1269	4.0896	0.0000
c6/MC	0.6285	0.3950	8.6666	0.0000
c7/MC	0.4936	0.2436	6.0871	0.0000
c8/MC	0.6622	0.4386	9.4792	0.0000

Table 6 Correlations between Industry 4.0/5.0 technologies implemented and the increase in the level of MC strategy in the automotive industry

Relations	Correlation	r^2	t	p
I1/MC	0.3859	0.1489	4.4870	0.0000
I2/MC	0.4739	0.2246	5.7715	0.0000
I3/MC	0.4153	0.1724	4.8961	0.0000
I4/MC	0.4739	0.2246	5.7715	0.0000
I5/MC	0.3362	0.1130	3.8285	0.0002
I6/MC	0.5438	0.2957	6.9500	0.0000
I7/MC	0.1867	0.0348	2.0390	0.0437
I8/MC	0.2137	0.0456	2.3464	0.0206
I9/MC	0.6285	0.3950	8.6666	0.0000
I10/MC	0.1393	0.0194	1.5086	0.1341
I11/MC	0.0976	0.0095	1.0522	0.2949
I12/MC	0.3957	0.1566	4.6218	0.0000
I13/MC	0.4739	0.2246	5.7715	0.0000

In addressing the third research question, the survey also investigated the benefits of implementing the MC strategy within the automotive industry. Among Polish SMEs, 36.75 % reported an increase in production flexibility, 33.33 % noted a reduction in waste, and 28.21 % observed an annual sales increase of up to 10 %. Additionally, 26.50 % of SMEs declared improvements in workforce flexibility and decentralization of decision-making. A reduction in electricity consumption was reported by 24.79 % of enterprises, while 20.51 % highlighted more effective analysis of large data sets. Subsequently, the relationship between these outcomes of MC strategy implementation and the increase in the level of MC strategy was analyzed (see Table 7).

The primary relationships identified in Polish SMEs between the implementation of the MC strategy and increased production flexibility (0.5644), as well as waste reduction (0.5235), were found to be significant.

Table 7 Correlations between the outcomes of MC strategy implementation and the increase in the level of MC strategy in the automotive industry

Relations	Correlation	r^2	t	p
E1/MC	0.4641	0.2153	5.6188	0.0000
E2/MC	0.2729	0.0745	3.0431	0.0029
E3/MC	0.1564	0.0244	1.6986	0.0920
E4/MC	0.4445	0.1976	5.3223	0.0000
E5/MC	0.5644	0.3186	7.3328	0.0000
E6/MC	0.5235	0.2741	6.5904	0.0000
E7/MC	0.4250	0.1806	5.0360	0.0000
E8/MC	0.3761	0.1414	4.3535	0.0000
E9/MC	0.4445	0.1976	5.3223	0.0000

4. Discussion

The research findings provide comprehensive answers to the three posed research questions. The primary factors driving the enhancement of the Mass Customization (MC) strategy implementation level in Polish manufacturing SMEs within the automotive sector include active customer involvement in both the MC product design process and the utilization of real-time configuration tools tailored to meet customer requirements. A notable example of such an approach is the Customer-Product Interaction Life Cycle (CILC) model [26], which facilitates cost reduction and enhances customer satisfaction. Furthermore, the results demonstrate that the adoption of Industry 4.0/5.0 (I4.0/5.0) technologies, particularly automated machinery, contributes significantly to the advancement of MC strategy implementation.

The study also delineates future research directions in MC, emphasizing: (1) the development of enterprise strategies that integrate customer participation in new product and process design; (2) the incorporation of automation and robotics in alignment with MC objectives; and (3) the transformative impact of I4.0/5.0 technology adoption on MC practices.

Nevertheless, certain limitations warrant attention and suggest avenues for further investigation. First, the current study is confined to the automotive industry, focusing on customized orders. Extending this research to other industrial sectors, such as metal manufacturing, could reveal whether observed patterns are generalizable or industry-specific. Second, the study does not address manufacturer performance metrics. Financial constraints typical of SMEs restrict the deployment of I4.0/5.0 technologies, thereby limiting MC strategy realization in production processes. The substantial investment required for implementing I4.0/5.0 solutions represents a significant barrier to enhancing MC production capabilities [27]. Future research should, therefore, integrate analyses of investment expenditures with assessments of both tangible and intangible benefits derived from these technologies, as such transformations are imperative for sustaining competitive advantage.

Third, the present investigation centers on SMEs, which constitute the majority (over 90 %) of enterprises in Poland. Subsequent research should explore the interplay between Industry 4.0/5.0 adoption (i.e., acquisition and utilization of advanced technologies) and the simultaneous increase in customer satisfaction and profit margins across small, medium-sized, and large enterprises. It is crucial to recognize that these categories possess distinct financial strategies and capital limitations.

This article represents the initial phase of a broader research initiative on MC strategy implementation in manufacturing enterprises. Building on the collected data, an AI-based predictive model will be developed to validate the empirical findings and uncover potential additional or non-linear relationships undetectable through traditional correlation analyses [28]. The model will employ artificial neural networks trained and tested on survey data filtered by significance analysis. The subsequent research stage will leverage this model to simulate and identify strategies that optimize the desired level of product customization.

5. Conclusion

This paper presents a diagnostic analysis of the situation regarding Mass Customization (MC) realization on the example of Poland in the European market. In enterprises across European Union countries, notable similarities can be observed in the functioning of the market, partly due to the regulations of the European Commission. Therefore, our research results in MC can be treated as a reference framework for activities aimed at enhancing the MC level in European manufacturing enterprises. We hope that the results of our research will inspire further studies to explore these areas within other countries with varying geopolitical and economic conditions. Recommendations for managers in the automotive manufacturing sector were formulated regarding necessary actions to increase the level of MC strategy. Firstly, it is advised to implement tools that will enable customers to participate in the design of the product and process, and secondly, to develop mechanisms that will enable the analysis of real-time data and the implementation of automatic solutions, even when dealing with individual projects. This is undoubtedly a challenge and a promising avenue for further research in the field of improving MC strategy implementation.

Acknowledgments

This work was supported by a program of the Polish Ministry of Science under the title ‘Regional Excellence Initiative’, project no. RID/SP/0050/2024/1.

Declaration of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Fathi, M., Ghobakhloo, A. (2020). Enabling mass customization and manufacturing sustainability in Industry 4.0 context: A novel heuristic algorithm for in-plant material supply optimization, *Sustainability,* Vol. 12, No. 16, Article No. 6669, doi: 10.3390/su12166669.

[2] Kim, S., Lee, K. (2023). The paradigm shift of mass customisation research, *International Journal of Production Research,* Vol. 61, No. 10, 3350-3376, doi: 10.1080/00207543.2022.2081629.

[3] Patalas-Maliszewska, J., Kowalczewska, K., Rehm, M. (2023). The role of additive manufacturing in supply chain management, *Management and Production Engineering Review,* Vol. 14, No. 1, 25-33, doi: 10.24425/mper.2023.145363.

[4] Bojic, S., Maslaric, M., Mircetic, D., Nikolicic, S., Todorovic, V. (2023). Simulation and Genetic Algorithm-based approach for multi-objective optimization of production planning: A case study in industry, *Advances in Production Engineering & Management,* Vol. 18, No. 2, 250-262, doi: 10.14743/apem2023.2.471.

[5] Suzic, N., Forza, C. (2023). Development of mass customization implementation guidelines for small and medium enterprises (SMEs), *Production Planning & Control*, Vol. 34, No. 6, 543-571, doi: 10.1080/09537287.2021.1940345.

[6] Song, O., Ni, Y., Ralescu, D.A. (2021). Product configuration using redundancy and standardisation in an uncertain environment, *International Journal of Production Research,* Vol. 59, No. 21, 6451-6470, doi: 10.1080/00207543.2020.1815888.

[7] Mönch, T., Huchzermeier, A., Bebersdorf, P. (2021). Variable takt times in mixed-model asmbly line balancing with random customization, *International Journal of Production Research,* Vol. 59, No. 15, 4670-4689, doi: 10.1080/00207543.2020.1769874.

[8] Ebrahimi, M., Mahmoodjanloo, M., Einabadi, B., Baboli, A., Rother, E. (2023). A mixed-model assembly line sequencing problem with parallel stations and walking workers: A case study in the automotive industry, *International Journal of Production Research*, Vol. 61, No. 3, 993-1012, doi: 10.1080/00207543.2021.2022801.

[9] Mourtzis, D., Angelopoulos, J., Panopoulos, N. (2022). Personalized PSS design optimization based on digital twin and extended reality, *Procedia CIRP*, Vol. 109, 389-394, doi: 10.1016/j.procir.2022.05.267.

[10] He, Y., Stecke, K.E. (2022). Simultaneous part input sequencing and robot scheduling for mass customization, *International Journal of Production Research*, Vol. 60, No. 8, 2481-2496, doi: 10.1080/00207543.2021.1894369.

[11] Wang, Y., Zhao, W., Wan, W.X. (2021). Needs-based product configurator design for mass customization using hierarchical attention network, *IEEE Transactions on Automation Science and Engineering,* Vol. 18, No. 1, 195-204, doi: 10.1109/TASE.2019.2957136.

[12] Vještica, M., Dimitrieski, V., Pisarić, M.M., Kordić, S., Ristić, S., Luković, I. (2023). Production processes modelling within digital product manufacturing in the context of Industry 4.0, *International Journal of Production Research*, Vol. 61, No. 19, 6271-6290, doi: 10.1080/00207543.2022.2125593.

[13] Pech, M., Vrchota, J. (2022). The product customization process in relation to Industry 4.0 and digitalization*, Processes*, Vol. 10, No. 3, Article No. 539, doi: 10.3390/pr10030539.

[14] Bouchard, S., Gamache, S., Abdulnour, G. (2023). Operationalizing mass customization in manufacturing SMEs – A systematic literature review, *Sustainability*, Vol. 15, No. 4, Article No. 3028, doi: 10.3390/su15043028.

[15] Yang, H., Jiang, L. (2024). Modular design of new products and remanufacturing of used products under mass customization, *RAIRO-Operations Research*, Vol. 58, No. 1, 103-128, doi: 10.1051/ro/2023151.

[16] Zeng, F. (2024). Research on key technology of mass customization based on flexible manufacturing in the context of deep learning, *Applied Mathematics and Nonlinear Sciences*, Vol. 9, No. 1, 1-21, doi: 10.2478/amns-2024-0115.

[17] Nair, V.V., Kuhn, D., Hummel, V. (2019). Development of an easy teaching and simulation solution for an autonomous mobile robot system, *Procedia Manufacturing,* Vol. 31, 270-276, doi: 10.1016/j.promfg.2019.03.043.

[18] Tilbury, D.M. (2019). Cyber-physical manufacturing systems, *Annual Review of Control, Robotics, and Autonomous Systems,* Vol. 2, 427-443, doi: 10.1146/annurev-control-053018-023652.

[19] Dou, Z., Sun, Y., Wu, Z., Wang, T., Fan, S., Zhang, Y. (2021). The architecture of mass customization-social Internet of Things system: Current research profile, *ISPRS International Journal of Geo-Information,* Vol. 10, No. 10, Article No. 653, doi: 10.3390/ijgi10100653.

[20] Zhang, Z., Wang, X., Zhu, X., Cao, Q., Tao, F. (2019). Cloud manufacturing paradigm with ubiquitous robotic system for product customization, *Robotics and Computer-Integrated Manufacturing*, Vol 60, 12-22, doi: 10.1016/j.rcim.2019.05.015.

[21] Burge, T.A., Jeffers, J.R.T., Myant, C.W. (2022). Development of an automated mass-customization pipeline for knee replacement surgery using biplanar X-rays, *Journal of Mechanical Design*, Vol. 144, No. 2, Article No. 021702, doi: 10.1115/1.4052192.

[22] Sarker, I.H., Kayes, A.S.M., Badsha, S., Alqahtani, H., Watters, P., Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective, *Journal of Big Data*, Vol. 7, Article No. 41, doi: 10.1186/s40537-020-00318-5.

[23] Wang, Y., Ma, H.-S., Yang, J.-H., Wang, K.-S. (2017). Industry 4.0: A way from mass customization to mass personalization production, *Advances in Manufacturing*, Vol. 5, 311-320, doi: 10.1007/s40436-017-0204-7.

[24] Mankins, J.C. (2009). Technology readiness assessments: A retrospective, *Acta Astronautica,* Vol. 65, No. 9-10, 1216-1223, doi: 10.1016/j.actaastro.2009.03.058.

[25] Koppenhagen, F., Held, T. (2021). The implications of product modularisation on the development process, supplier integration and supply chain design in collaborative product development, *Advances in Production Engineering & Management*, Vol. 16, No. 1, 82-98, doi: 10.14743/apem2021.1.386.
[26] Zhang, X., Ming, X. (2024). A smart system of customer- product interaction life cycle (CILC) in industrial internet era for mass personalization from industrial practice survey: Identification, definition, acquisition and parsing, *Journal of Intelligent Manufacturing,* Vol. 35, 727-756, doi: 10.1007/s10845-022-02070-4.
[27] Alqoud, A., Schaefer, D., Milisavljevic-Syed, J. (2022). Industry 4.0: challenges and opportunities of digitalisation manufacturing systems, In: *Proceedings of the 19th International Conference on Manufacturing Research, Incorporating the 36th National Conference on Manufacturing Research*, Derby, UK, 25-30, doi: 10.3233/ATDE220560.
[28] Patalas-Maliszewska, J., Pająk, I., Skrzeszewska, M. (2020). AI-based decision-making model for the development of a manufacturing company in the context of Industry 4.0, In: *Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE*), Glasgow, United Kingdom, 1-7, doi: 10.1109/FUZZ48607.2020.9177749.

APEM journal

Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 29–42
https://doi.org/10.14743/apem2025.1.525

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Closed-loop simulation of workload control: Integrating input-output regulation with feedback

Oliveira, A.F.[a], Sagawa, J.K.[a], Mušič, G.[b,*]

[a]Federal University of São Carlos, Department of Production Engineering, Rodovia Washington Luís, São Carlos, Brazil
[b]University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

A B S T R A C T

When responsiveness to demand fluctuations is a key performance factor, continuous dynamic models can advantageously replace discrete event simulation, often employed in Workload Control (WLC) studies. Nonetheless, dynamic modelling tools and feedback control are scarcely applied to WLC or other shop floor control systems. To fill this gap, this paper presents a closed-loop model of WLC incorporating feedback control and shop floor input-output control. The bond graphs' dynamic modelling technique is employed. The model is implemented in Simulink® and its behaviour in face of disturbances is analysed. The load of the machines and of the job pool of WLC is considered to adjust order release (input control). The capacity of each machine (output control) is altered in function of the level of its preceding buffer. The machines' processing rates stabilize and the reference levels for the buffers are reached when a step disturbance in order entry is simulated. Also, the system responded with a maximum capacity increment of 15 % when cyclic demand is simulated. This novel approach of a smart production control system can help managers to better control shop floor load in response to disturbances.

A R T I C L E   I N F O

Keywords:
Workload control;
Closed-loop systems;
Simulation;
Smart production systems;
Control theory;
Bond graphs;
Input-output control;
Shop floor

**Corresponding author:*
gasper.music@fe.uni-lj.si
(Mušič, G.)

Article history:
Received 14 January 2024
Revised 29 March 2025
Accepted 31 March 2025

1. Introduction

The Production Planning and Control (PPC) can be seen as a determinant factor to the performance of an organization, thus, the choice of adequate PPC Systems remains as a key point. Some classic PPC approaches cannot be well applied to production systems that are not based on demand forecasts, like make-to-order (MTO) systems, due to the fact that these approaches focus on large systems within a repetitive production environment [1]. The Workload Control (WLC) is an alternative to MTO systems with a job shop production flow. It balances the load of the workstations, allowing only controlled releases to the shop floor, reducing the variability in the order entry level [2]. The term load refers to the capacity required in a production system.

The production in job shops lead to queues of orders/work-in-process (WIP) competing for the workstation's capacity. The Workload Control concept seeks to achieve small and stable queues, i.e., to establish low and stable direct load levels [3].

The Workload Control rules apply to the discrete domain, i.e. to systems with discrete production orders. Most studies in the literature model and test this approach by means of discrete event simulation, with order arrival, due dates and processing times defined by probability distributions [4]. Other approaches from the dynamic modelling field are a relevant alternative to discrete event simulation (DES), since they allow a more analytical understanding of the fundamental dynamics of the system allow to automatically adjust WLC parameters in response to dynamic changes and other disturbances. While the discrete event simulation is a trial-and-error method [5], if used alone, the dynamic modelling allows the use of feedback loops, i.e., automatic control, generating smart closed loop models, i.e. models that yield prescriptive guidelines for WLC operation.

Different approaches for dynamic modelling and control can be used, such as transfer functions, difference equations, optimal control, etc. In [6], the author uses bond graphs (BG) and proposes the representation of production stations based upon analogies with electrical components like resistor, capacitor and source of effort. This representation can also be found in [7], and can be applied to Workload Control simulation.

This paper aims to model a Workload Control system as a continuous dynamic system with feedback loop (i.e., a closed-loop system), and analyse the behaviour of this system in face of demand disturbances. The proposed model is smart in the sense that the WLC parameters are automatically changed in function of the current state of the shop floor; this parameterisation is responsive to disturbances. In the extant WLC models, the parameterisation is done by means of a set of experiments with open-loop systems (the DES systems proposed in the extant literature are open-loop systems, i.e., systems without feedback loop).

The proposition of a closed-loop model (automatically controlled by a controller) is an unusual approach for WLC simulations. It provides intelligence to the production planning system. Currently, smart production planning and scheduling [8, 9] are relevant research trends. An additional gap in the Workload Control literature refers to research that deals, simultaneously, with input and output control. Input control in WLC is about restraining order release, while output control corresponds to adjustments in the capacity of the machines. The vast majority of WLC studies focus on input control; publications related to output control are scarce, and even scarcer is the research related to the simultaneous application of both (e.g. [10]). The proposed model also addresses this gap by proposing control rules for the application of both input and output controls.

Thus, the novelty and contributions of this research are summarized as follows:

- A novel closed-loop continuous model for WLC simulation is proposed;
- Input and output control are simultaneously operating in the same model, expanding the extant literature;
- The input-output control adjustments are given by a single continuous simulation run (for a given scenario) in contrast to prior research where the usual approach is to determine the control adjustments by multiple experiments;
- The resulting capacity increments are gradual, based on continuous curves, in contrast to fixed-step increments proposed in the literature (see Section 2).

2. Literature review

A search for papers presenting any type of relationship between Control Theory and Workload Control was conducted as theoretical base for this research. No closed-loop models (i.e. models with a controller and a feedback loop) of WLC systems were found. In [11], capacity adjustments of the stations are performed in function of the shop floor current state; however, the model does not present a controller. Instead, pre-defined rules are used to perform the adjustments, with constant values as increments/decrements. A few WLC papers with multi-agent systems were found. Moreover, the review included papers with the simultaneous application of input and output control in WLC. These two groups of papers will be discussed as follows.

In [12], a system with 3 agents is proposed for WLC implementation: order entry agent (OEA), job agent (JA), and machine agent (MA). The due date settings and customers' enquiries are performed by the OEA. The JA defines the job routings and the MA allocates the jobs in each machine queue.

A similar approach is presented by [13] using 4 agents. The first agent (OEA) demands information of current shop status and communicates with the JRA, which operates the job release using a continuous aggregate loading (CAGG) mechanism: jobs are continuously released to the shop if the workload is below the norm. The JRA communicates with the RSA to calculate the current workload [13]. The RSA encompasses job agents (JA) and machine agents (MA), with the same functions as in [12]. The fourth agent, IFA, estimates the average lateness, used by the OEA for decision-making [13].

The authors in [14] propose a WLC dynamic model based upon product agents and planning agents. The job pool agent (JPA) keeps data regarding products waiting in the job pool. The shop floor assessment agent (SAA) collects shop floor data about orders already released, and computes the shop floor simulation. The results are sent to the JRA, together with the information from the pool. The JRA assesses the impact of each potential order release and makes continuous release decisions.

The discussed multi-agent systems present similarities. Both [13] and [14] use a job release agent (JRA) for the release decision-making, and the job routing and sequencing agent (RSA) has mostly the same responsibilities as the shop floor assessment agent (SAA). However, in [13], CAGG is the job release mechanism, while [14] use priority indexes, defined to each product agent, for order sequencing and release.

The second branch of this review contains the papers that discuss WLC approaches with simultaneous control of input (order release) and output (capacity adjustment).

The authors in [15] present a conceptual view of different approaches in WLC. Fig. 1a portrays the input control, where the released work quantity is based on the shop floor queues. This is the most common approach in workload control research. Fig. 1b shows output control; the released work quantity is uncontrolled and the work centre capacities are controlled with the aim of keeping the queues at a constant level. Figs. 1c and 1d show possibilities for simultaneous input/output control. In the approach shown in Fig. 1c, both order release and work centre capacities are controlled on the basis of shop floor load. In Fig. 1d, only the order release is controlled based on the shop floor load while the work centre capacities are controlled according to the pool level.

In [11], the authors analyse the effects of input-output control in a pure job shop. Urgent and non-urgent jobs are organised in the pool according to different rules. The release is based on the LUMS COR method, which is periodic but has a trigger against station starvation. This represents the input control of the system. The capacity adjustments (output control) are based on three parameters: the size of the increment (α), the load threshold that triggers the capacity adjustment (upper limit of utilisation, β) and the load reduction threshold that triggers the adjustment setting (γ). The authors perform simulations with 5 levels for α (0, 10, 20, 30 or 40 %), 3 levels for β (85, 90 and 95th percentile) and 3 levels for γ (0, 5 and 10 %).

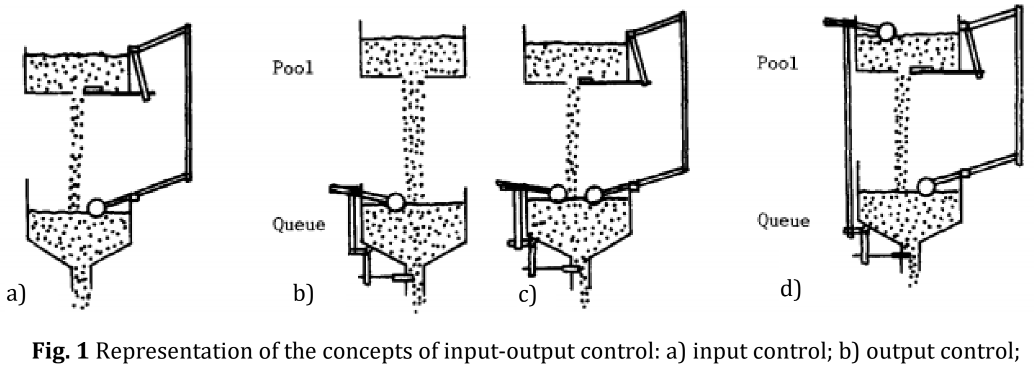


Fig. 1 Representation of the concepts of input-output control: a) input control; b) output control; c) and d): approaches for simultaneous input and output control [15]

The aforementioned study is based on discrete event simulation (DES), and constant capacity increments, chosen from a finite set of values, are applied. When utilization reaches a defined threshold (β), e.g. 85 %, a step-shaped increment (α) is applied; the extra capacity is deactivated after the utilization reduces a certain amount (γ). The representation of a WLC system as a continuous dynamic system (using bond graphs) allows the application of automatic controllers, yielding a closed-loop model. In this way, the WLC parameters, i.e. order release level and size of capacity increments, can be defined as continuous curves, based on the instantaneous shop floor load, instead of being defined as fixed/constant values based on experiments.

There is a paucity of research investigating the simultaneous application of WLC input and output control, as mentioned. The proposed closed-loop model and simulations serve as proof of concept for the alternatives presented in [15] and shown in Fig. 1, also complementing the (scant) existing input-output control research.

Considering a wider view, other advanced methods of production control were analysed in [16]. Model-based production control was proposed in combination with computational intelligence and data-based modelling methods. Among those, evolving model identification has been successfully applied to production systems with non-linear process behaviour [17] and combined with design of experiments [18]. A data-driven modelling approach has been successfully applied to energy consumption optimization in steel production [19].

The application of reinforcement learning (RL) to production control is also a recent research trend. Most models found in the literature that consider work-in-process (WIP) or workload control employ the CONWIP (Constant Work in Process) policy, such as [20-22], and not WLC. CONWIP is simpler than WLC, because only the total workload of the system matters, instead of the workload of each machine. In CONWIP, a new order or amount of work is released only when an order or amount of work outputs the system, keeping the total work-in-process constant. A fixed number of cards or containers is used to limit the WIP. In the aforementioned works, an agent that learns by reinforcement is coupled to a DES system representing the shop floor. The agent observes the current state of the system and modifies the amount of cards in it, i.e., it adjusts the total WIP allowed in the system dynamically, influencing the system's throughput. This production control policy is simpler than WLC or than other production control systems because there is just one parameter to be adjusted. No works specifically applying reinforcement learning to WLC were found.

A similar framework (i.e., combining DES and decision-making algorithms) is employed by [23] to a system ruled by the Make-To-Availability (MTA) policy. With MTA, for a given product, the orders are released and sequenced according to the difference between the established target level for the buffer of that product and the actual inventory on hand. The bigger this difference is, the higher is the priority of the order. In [23], the target level of the buffers is adjusted considering the current state of the plant, which is influenced by demand fluctuations and variability of the processing times, among others.

A literature search in Scopus database with the string *("artificial intelligence" OR "machine learning" OR "reinforcement learning") AND "workload control"* was carried out, returning 10 papers. The string was applied to the field "Title, Abstract and Keywords". Among these 10 papers: two were from other areas (i.e., used the term "workload" within other contexts); one [20] applied reinforcement learning to CONWIP, and not to WLC; six investigated specific parameters/aspects of WLC but not related to machine learning (ML) or artificial intelligence (AI), and one treated about WLC software development. This reinforces the existence of a relevant gap to be tackled, concerning the dynamic parameterisation of WLC and the incorporation of AI and ML techniques into WLC research.

3. Modelling and simulation

The system simulated in this paper was presented by [4]. Jobs and material are provided by the source $S_{01\text{-}pool}$, which is connected to the job pool, i.e. the queue of jobs/orders awaiting release. The shop floor has four workstations (Fig. 2), representing a general flow production [24]. This means that there are different product families with different production routings, without reentrant flows (the flow is directional). Earlier studies on output control, such as [11], considered pure job shops and not general flow shops. Using the data on product families, routings and demand (Table 1, taken from [4]), the percentage of material that must flow into each branch of the system can be calculated (Fig. 2).

Table 1 Production routings and demand rates for the unidirectional job shop

	Station 1	Station 2	Station 3	Station 4	Demand rate (units/day)
Family 1	X	X			1.5
Family 2	X	X		X	1.25
Family 3			X		1.3
Family 4			X	X	1.1

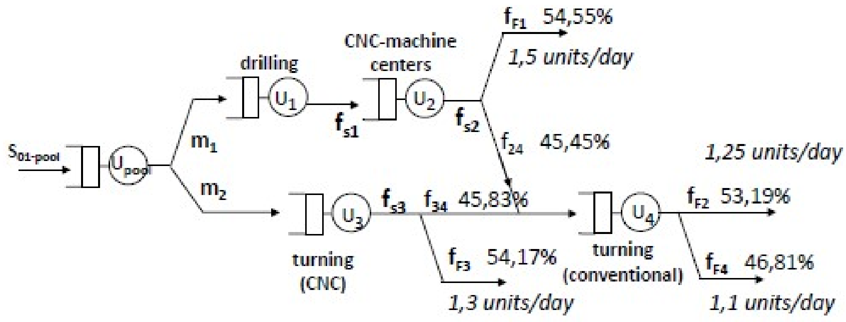


Fig. 2 Schematics of the modelled WLC system

A modelling approach based on bond graphs is chosen for the development of a mathematical model of the production system. Bond graphs capture the power and energy exchange between the model elements. The graphical representation of these elements and their interconnections can be translated into equations that form a mathematical model.

To represent a shop floor with bond graphs requires some adaptations of the standard bond graph formalism. A corresponding bond graph model for a production station, based on analogies to capacitor, resistor and source of effort, was proposed in [6]. The representation focuses on fluidised, continuous material flow (f), which can also be interpreted as the amount of work per unit of time. The approach in [6] assumes an unlimited buffer capacity, which allows the output flow of the i-th production station to be expressed as:

$$f_i = U_i \min(1, q_i), \tag{1}$$

where U_i is the processing frequency of the station (i.e. the reciprocal of a processing time) and q_i is the amount of material stored in the i-th buffer. These variables are dependent on time, e.g. $f(t)$, but the time dependence is not explicitly shown in the equations for simplicity. An analogy is assumed between the material flow rate of a machine and the electric current through a resistor, where U_i is the reciprocal of the resistance ($U_i = 1/R_i$). The expression $\min(1, q_i)$ stands for the effort (analogue to voltage), which gradually decreases with a small amount of material storage. Further details on this modelling approach can be found in [6].

The material quantity rate in the station buffer ($\dot{q}_i$) depends on the balance between the material input rate and material consumption rate, which is expressed by the difference between material input and material output flows of the station. This leads to equation:

$$\dot{q}_i = f_{ei} - U_i \min(1, q_i), \tag{2}$$

where f_{ei} is the material input flow of station i while U_i and q_i have the same meaning as in (1). The use of these equations for modelling multiproduct manufacturing systems is presented in [7, 25, 26]. In the simulated model, the controlled variables are the levels of the buffers ($q_i, i = 1,2,3,4$) and of the job pool (q_{pool}), and the manipulated variables are the processing frequencies of the machines (U_i, with $i = 1,2,3,4$), and the release frequency of the job pool (U_{pool}). The most important variables are summarised here:

- $\dot{q}_{pool}$: change rate of waiting orders/materials in the pool;
- $\dot{q}_i$: rate of change of the material quantity in buffer *i*;
- q_{pool}: quantity of orders waiting to be released;
- q_i : quantity of material in machine buffer *i*, where *i* = 1,2,3,4;
- U_{01pool}: rate/frequency of orders/materials entering the job pool;
- U_{pool}: release frequency. In the model, the release is considered an operation and the pool operates similarly to a work centre;
- U_i : processing frequency of machine *i*, with *i* = 1,2,3,4.

The products' demand rates (Table 1), which can be modified or modelled as curves to represent volatile demand, are the main parameters of the system. The model is also parameterized by m_1 and m_2, which will be described later.

Applying (2) to the pool and buffer of station 2 gives:

$$\dot{q}_{pool} = U_{01pool} - U_{pool}\min(1, q_{pool}) \tag{3}$$

$$\dot{q}_2 = U_1 \min(1, q_1) - U_2 \min(1, q_2). \tag{4}$$

Stations 1 and 2 are connected in series, so that the input of station 2 in (4) is replaced by the output of station 1.

Convergent and divergent junctions conduct parts of the flow to different stations, according to the production routings shown in Fig. 2. Convergent junctions establish the flow conservation and are represented as a 1-junction in bond graphs, leading to:

$$f_{si} = \sum_{p=1}^{P} f_{ep}, \tag{5}$$

in which f_{si} represents the output flow of a *i*-th convergent junction, and f_{ep}, *p* = 1,2, ..., *P*, represent the branches of incoming flow. The divergent junctions can be represented as transformers in bond graphs. The q-th transformer with module m_q receives an input flow f_{ei} and converts it into an output flow f_{sq}:

$$f_{sq} = m_q f_{ei}, \tag{6}$$

with $\sum_q^Q m_q = 1$ and *q* = 1,2, ..., *Q*, where Q is the total number of diverging branches [4].

The proposed model has modules m_1 and m_2, which describe how the material flow is distributed between the two processing branches coming from the pool (see the first junction of Fig. 2). The application of (6) to this junction and of (2) to station 1 and station 3 yields:

$$\dot{q}_1 = m_1 U_{pool}\min(1, q_{pool}) - U_1 \min(1, q_1) \tag{7}$$

$$\dot{q}_3 = m_2 U_{pool}\min(1, q_{pool}) - U_3 \min(1, q_3). \tag{8}$$

In the case considered, the parameters were set to m_1 = 0.53398 and m_2 = 0.46602 to obtain the desired production mix. The application of (5) and of (6) considering the percentages of split shown in Fig. 2 yields:

$$f_{e4} = f_{24} + f_{34} \tag{9}$$

$$f_{24} = 0.4545 U_2 \min(1, q_2) \tag{10}$$

$$f_{34} = 0.4583 U_3 \min(1, q_3) \tag{11}$$

$$\dot{q}_4 = 0.4545 U_2 \min(1, q_2) + 0.4545 U_3 \min(1, q_3) - U_4 \min(1, q_4). \tag{12}$$

The instantaneous WIP in the shop floor (q_i) is calculated by the integral of the rates of material storage or consumption in the buffers, and the amount of orders waiting in the pool (q_{pool}) is obtained by the integral of the release rate of orders/material in the pool. These are the state variables of the system (q_i, q_{pool}). The algebraic manipulation of (3), (4), (7), (8), and (12) yields the state model:

$$\begin{bmatrix}\dot{q}_{pool}\\ \dot{q}_1\\ \dot{q}_2\\ \dot{q}_3\\ \dot{q}_4\end{bmatrix}=\begin{bmatrix}-1 & 0 & 0 & 0 & 0 & 1\\ m_1 & -1 & 0 & 0 & 0 & 0\\ 0 & 1 & -1 & 0 & 0 & 0\\ m_2 & 0 & 0 & -1 & 0 & 0\\ 0 & 0 & 0.4545 & 0.4583 & -1 & 0\end{bmatrix}\begin{bmatrix}U_{pool}min\,(1,q_{pool})\\ U_1 min\,(1,q_1)\\ U_2 min\,(1,q_2)\\ U_3 min\,(1,q_3)\\ U_4 min\,(1,q_4)\\ U_{01pool}\end{bmatrix} \quad (13)$$

Details about this modelling, the development of the equations, and the variables and parameters can be obtained in [4, 25, 26].

In WLC, input control refers to the release of orders, which corresponds to the control of the processing frequency of the job pool, U_{pool}. Output control refers to the control of the processing frequencies of each machine, U_1 to U_4, which correspond to capacity adjustments. These controls are simultaneously executed in the proposed model. Release decisions are made in function of the shop floor load and the pool's load. Since all the production routings start either in station 1 or station 3 (Fig. 2), the buffer level of these stations is considered as a measure of the shop floor load. The load of the pool is measured based on the level of its own buffer of orders, q_{pool}. The level of the buffer upstream the *i*-th machine is assessed to define the capacity adjustment of the respective machine. As an analogy to Fig. 1, Fig. 3 represents the input and output control approaches used in all the simulations.

The approach adopted is similar to the one of Fig. 1c, with the releases based on the shop floor load and the capacity adjustments based on the machines' preceding buffers. However, in the proposed model, the releases also rely on the job pool load, differing from [15] and from other WLC studies. Moreover, in none of the works reviewed in Section 2, input and output control of Fig. 1c is implemented as a dynamic closed-loop system. To implement this system, a controller is attributed to each machine and to the pool. Fig. 4 shows the information that is used by each controller to adjust the processing frequencies U_{pool} and U_i.

Three different simulations were carried out. In all of them, the reference values for the buffer levels are: q_{pool_c} = q_{1c} = q_{2c} = q_{3c} = 7 and q_{4c} = 6. In simulation A, the system starts with empty buffers, to verify if stable operation would be achieved. Simulations B and C aimed to verify the system's behaviour in face of disturbances in the order entry level (which represents external demand). In both simulations, the buffer levels (WIP) start at the reference values, to check how the disturbances impact a system initially in balance. For Simulation B a step signal was chosen to represent the disturbances, and for Simulation C, a sinusoidal signal. Reference processing frequencies (U_{ip}) for the machines and for the sources were calculated based on (13) and on the demand rates shown in Fig. 2. These processing frequencies (U_{01pool_p} = 5.15, U_{pool_p} = 5.15, U_{1p} = 2.75, U_{2p} = 2.75, U_{3p} = 2.4, U_{4p} = 2.35) lead to the attendance of the demand in the medium term (in steady state) and are used as parameters for the control of the machines and the pool.

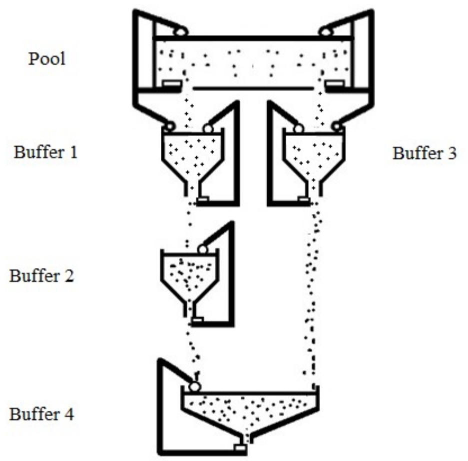


Fig. 3 Representation of simultaneous input and output control applied (analogy with Fig. 1)

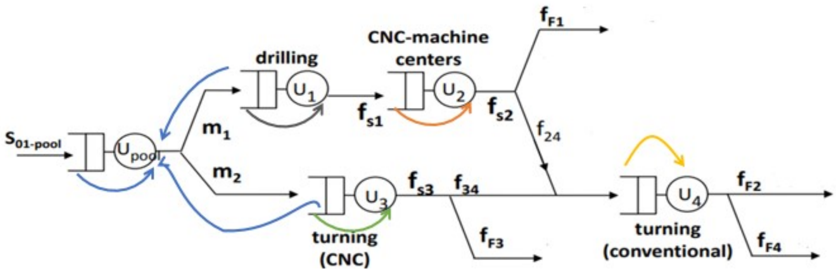


Fig. 4 signals sent to the controllers to implement the control rules (right side)

4. Results and discussions

4.1 Implementation

The model was implemented in Matlab and Simulink® (Fig. 5). The blue box represents the proportional integral (PI) controllers added to the system, which will define the processing frequencies of the pool and of each machine (U_{pool} and U_i). These instantaneous processing frequencies fluctuate around the steady state processing frequencies presented at the end of Section 3 (i.e. the output of the controllers are used as relative values to modify the steady state frequencies). The components highlighted in green represent the arrival of data for the controllers. The pool's controller receives 3 signals: the error from its own buffer, and the error from buffer levels 1 and 3. The controller of each machine receives only error signal of its preceding buffer. This implementation is aligned to the proposed control rules (Fig. 4).

The control signals of the system will be inputs for the production function, represented by the yellow box. This function calculates the rates of production or consumption of material ($\dot{q}_{pool}$ and $\dot{q}_i$) according to the state equation (13). In this block, there is also an integrator to convert the rates into work in process (WIP) stored in the pool and in the buffers. These values are, then, fed back into the system – as represented by the purple lines – and compared with the reference values in (red box). This comparison defines the error, which will serve as input for the controllers, closing the loop.

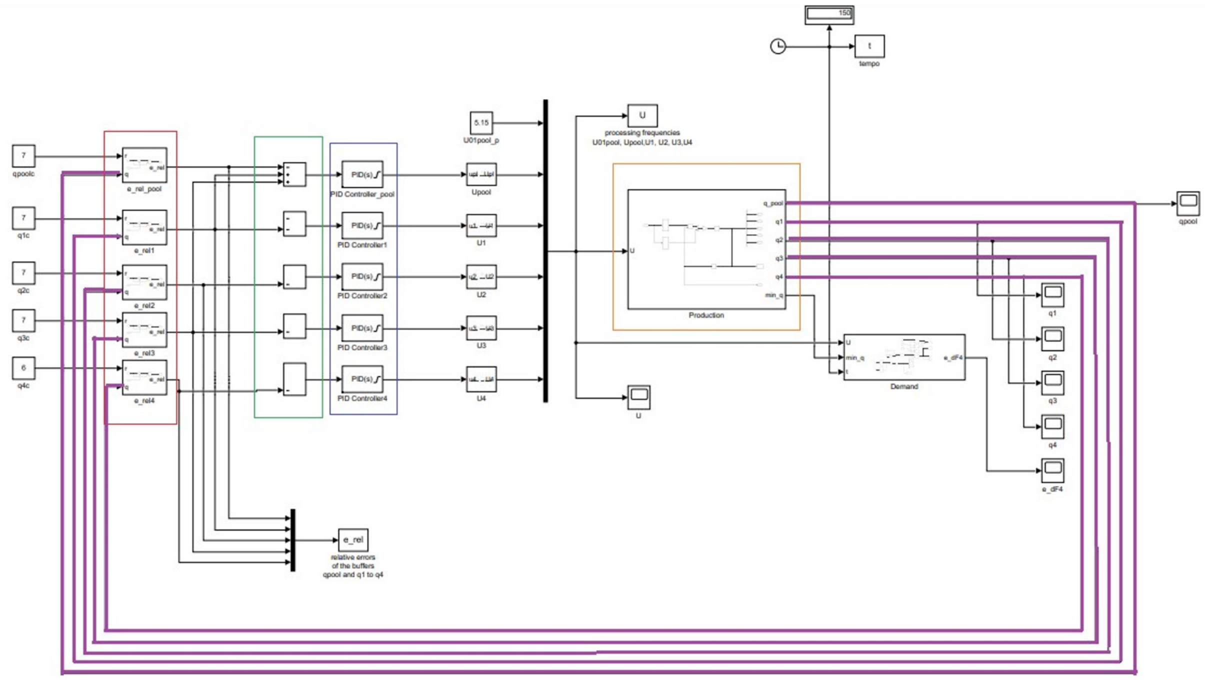


Fig. 5 Simulink® representation of the proposed model.

4.2 Results

The results of simulation A are shown in Figs. 6a and 7a. The controller of the pool receives signals from the buffers 1 and 3 that lead to the increase of the order release frequency, since the levels of both buffers are below the reference. A signal in the opposite direction comes from the pool itself. Since the level of orders in the pool is also below the reference, it induces a reduction in the release frequency, to clog the orders. The signals coming from the buffers 1 and 3 have a stronger influence on the control at the initial moments, and the order release frequency (U_pool) increases: the pool release frequency starts at 200 % of the release frequency in the steady state. This value decreases gradually as the levels of the buffers 1 and 3 increase and get closer to the reference (Fig. 7a). The inflection point of the release frequency curve occurs when machines 2 and 4 start processing, stimulating an increase in the flow through the whole system (i.e. the release frequency starts to increase again). After the transient period, the WIP stabilize at the reference values, showing that the control rules are effective.

In simulations B and C, the initial buffer levels match the reference levels. In Simulation B, there is a 20 % step-shaped increase in the order's arrival until the sixtieth day. In response to that, the release frequency of the pool increases due to the high levels of orders on wait. Next, there is an increase in the processing frequencies of the machines immediately downstream (machines 1 and 3). Finally, the reaction reaches the machines 2 and 4 (Fig. 6b).

The variations of the work in process in the buffers is not accentuated, showing that even with the disturbances – almost 20 % increase on the entry signal – the system did not suffer great penalties related to the buffers' levels (Fig. 7b). This behaviour results from the coordinated increase in the release frequency and in the processing frequencies of the machines (capacity adjustments).

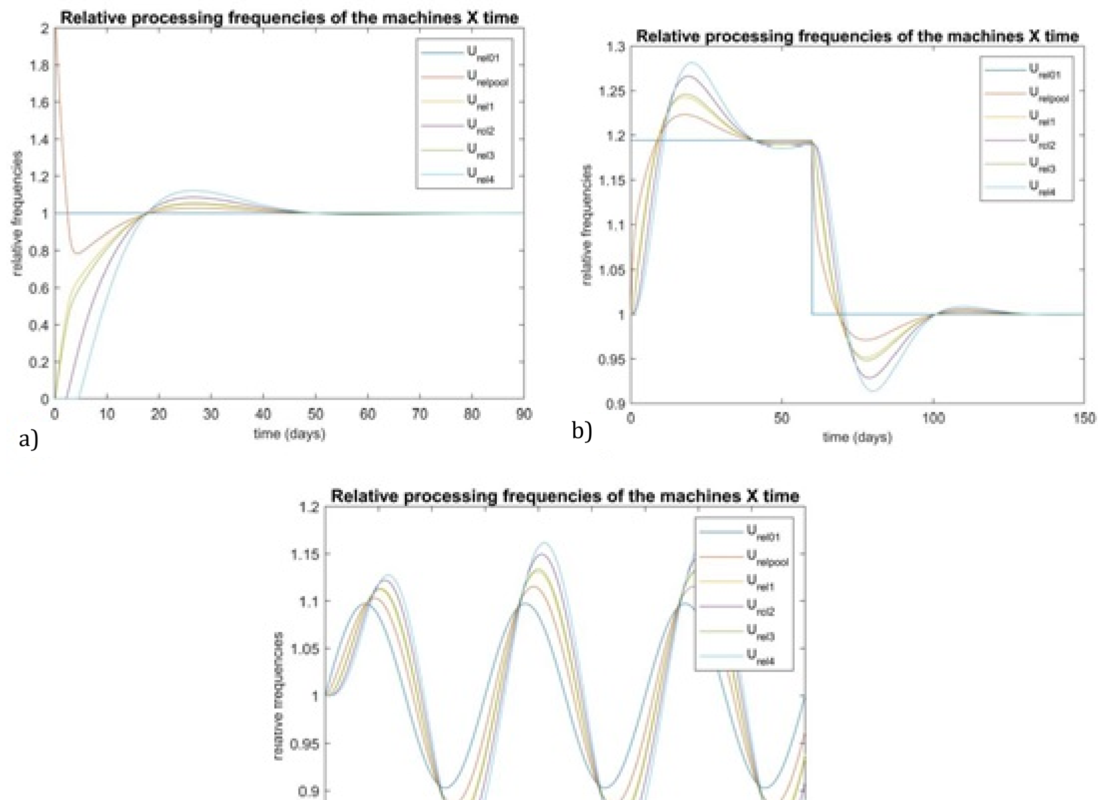


Fig. 6 Relative processing frequencies of the release and of the machines for:
a) Simulation A; b)Simulation B; c) Simulation C

With the return to the standard demand/input values (sixtieth day), the processing frequencies decrease and there is an undershoot before the system returns to operate at its regular capacity. This also reflects in an undershoot in the buffer levels. After that, they are led to the reference values again.

In Simulation C, the entry signal presents a sinusoidal variation of 10 % of the nominal value during the simulation (see curve of U_{rel01} in Fig. 6c).

The system regulates itself. The reaction begins with a higher release in the pool, and more material flows downstream with the increase of the processing capacity of the machines. A small delay in the reactions can be seen. The amplitude of the oscillations also gradually increases downstream. The last machine of the routings (Machine 4) has the higher peak and valley, with amplitude of almost 15 % of the nominal value. Meanwhile, the pool, the element most upstream, has the smallest variation, of 10 % (Fig. 6c).

The general behaviour of the system follows the behaviour of the entry signal, with a certain delay and with some amplification of the disturbance. The downstream stations are more affected. The system responds to the cyclic demand variations with a maximum 15 % of capacity variation (Fig. 6c).

The system reaches the steady state and the reference values for the buffers are achieved (simulations A and B). In simulation C, even with an increase of almost 20 % in the entry signal (demand signal), the WIP does not variate excessively (Fig. 7c). These are positive results, considering that the goal of WLC is to obtain a more stable and predictable shop floor. From the managerial point of view, it means that the system reacted to the demand increase without a strong increase the throughput time. This was only possible due to the increase in the processing frequencies (capacity adjustments). Managers, however, must analyse the viability of implementing the suggested increases in the machine's capacity.

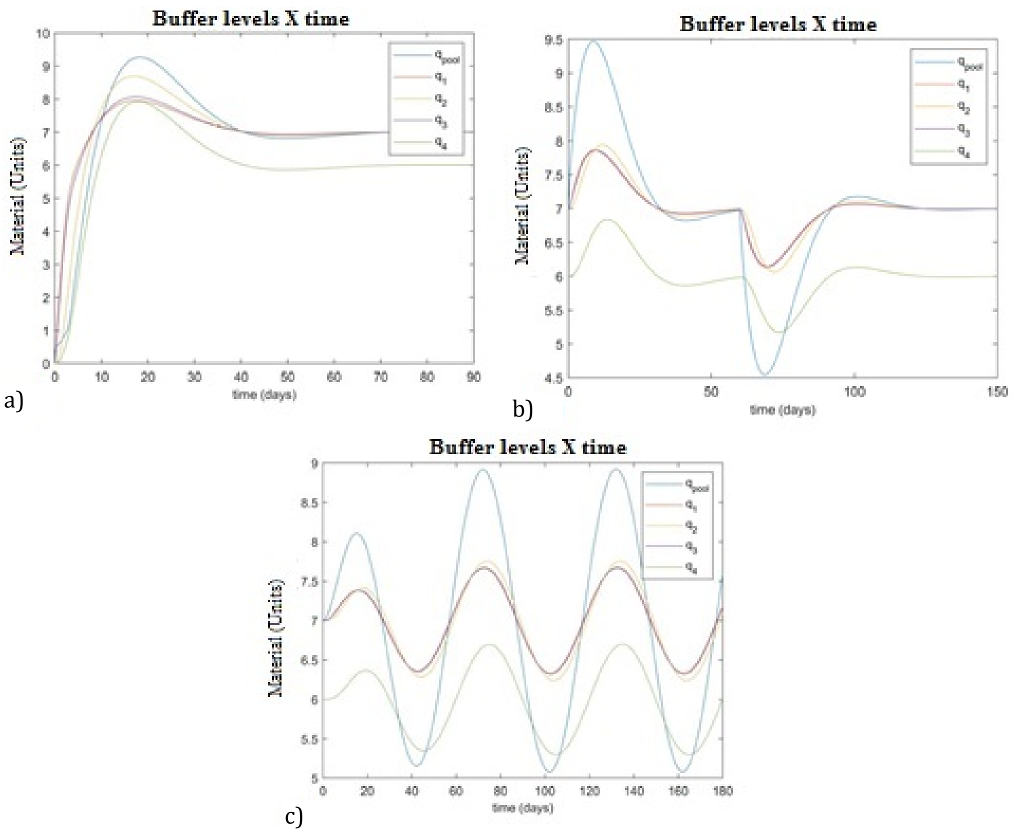


Fig. 7 Buffer levels for: a) Simulation A; b) Simulation B; c) Simulation C

The model developed presents a prescriptive characteristic. The results of simulation of the closed-loop model (i.e., with an automatic controller) point out when and with which amount adjustment actions must be taken to attend the demand fluctuations. Based on that, managers can define actions to implement the capacity increase or decrease, such as the reallocation of personnel, application of extra shifts, or the reduction of working hours of determined resources (i.e. defining programmed pauses in the schedule of the machines), when demand falls.

4.3 Discussions and model extensions

The applicability of the model in practice as well as extensions for its application to real and more complex shop floors and to supply chains are discussed as follows.

In the real manufacturing environment, the model could be implemented relatively easily from a technical point of view if the manufacturing environment is already supported by information systems, e.g. a Manufacturing Execution System (MES). Such a system collects data that can be used to align model states with the real environment and also offers options for adjusting capacities. More challenging is the underlying organisational change that would redefine the established concepts of production control. Transition to Industry 4.0 brings increasing demands of digitalisation of production processes in smart factories, e.g. through the use of Digital Twins and digital agents [27] in production control-related decision-making processes. This increases also the opportunities for incorporating advanced feedback control concepts such as the one proposed here.

Capacity adjustments can be implemented, in practice, in different ways. Small increments are usually implemented by means of overtime; substantial increments require long term managerial actions, as extra shifts, subcontracting or plant expansion. In the simulated scenarios, the capacity adjustments do not exceed 30 % (i.e. relative processing frequencies of 1.3 times the usual frequency, as shown in Fig. 6). Hence, they can be implemented by means of overtime and are associated only to variable costs, that is, it is supposed that extra fixed costs are negligible. Then, if a constant line $y = 1$ is drawn in the graphs of the relative processing frequencies, the cost of capacity adjustments can be related to the areas limited by the curves of processing frequencies and this reference line. For capacity increments, the areas above the line can be multiplied by a unitary cost of 1.5, if the overtime costs 50 % more than the normal work hour. Capacity decrements (or machines' downtime) also represent costs, because there are regular fixed costs related to the overall structure even when the machine is idle. So, the areas below the line $y = 1$ can be multiplied by a unitary cost (e.g., of 0.7) to estimate the cost of capacity decrements. These costs of gradual/continuous capacity adjustments can be compared to the costs of other policies, based on fixed adjustments, as the policy proposed by [11].

The advantage of dynamic models compared to discrete event simulation (DES) as a decision-making aid lies in the significantly lower model complexity. The probabilistic nature of DES models requires lengthy simulations before appropriate conclusions can be drawn. The dynamic model represents an aggregation of the complex DES dynamics. This comes at the cost of lower model accuracy, but the reduced complexity enables very fast simulation and even optimisation in real time, e.g. in the context of model predictive control. The lower model accuracy is compensated by the robustness of the feedback loop with the correctly tuned controller. Furthermore, aggregated dynamic models can also be used much more efficiently for sensitivity analyses and parameter calibration compared to DES. In contrast, DES models offer more flexibility and accuracy in simulating real emergency scenarios, such as machine failures and emergencies on the supply and/or demand side.

The model can be scaled to more complex shop floor configurations due to the modularity of the bond graphs' modelling technique. There is a bond graph model to a production station, with corresponding constitutive equations, derived from bond graphs' ideal elements, and there is a unique/fixed correspondence between an element of BG and its mathematical equation. The stations are linked by means of 1-junctions or transformer elements, as presented in Section 3. These elements allow representing different production routings, using different machine combinations. Thus, to represent shop floor configurations with much more machines and intricate flow, the modeller should add the machines and junctions into the bond graph pictorial representation, and then add the equations, corresponding to each element, into the mathematical model. After that, adequate algebraic manipulation leads to the matrix representation of the system. There is software, as 20-sim [28], that can automatically generate the mathematical/simulation model when the user elaborates the BG model by selecting the pictorial elements from a specific library. A real system in the textile industry with 11 machines and 9 different product routings was modelled by [7], showing that it is possible to scale the model. In the cited work, however, the shop floor control does not follow the principles of Workload Control. Since the goal is not to find an optimal solution to the models, but to simulate them with feedback control, the curse of dimensionality does not apply, i.e. the models are tractable and it is possible to simulate models with big dimensions using the computational capacity currently available.

External factors such as supply chain disruptions can affect the supply of raw materials to the system. The effect of these factors could be observed by simulating the temporary interruption of the flow coming from the source $S_{01\text{-}pool}$; i.e., the lack of raw material would prevent the entry of orders on the shop floor. If the suppliers take a significant time to recover from the disruption, the situation is similar to the one of simulation A, where the system has to start from the beginning, with empty buffers. The reference levels for the buffers is reached but with an overshoot, as seen in Fig. 7a. Another way to simulate supply chain disruptions would be to explicitly represent suppliers and customers as production stations linked to the plant already modelled. A delay in the output flow of the supplier's station can be added to represent the transportation and logistics times between the links; a delay in the output flow of the modelled plant can be also added to represent the logistics times needed to deliver the goods to the next or final customer. Other dynamic modelling tools such as block diagrams and transfer functions have been widely used to study the supply chain dynamics. For a review of these models, [29] can be consulted.

5. Conclusion

Simulations of WLC in the literature are based on open-loop discrete models. An alternative approach is presented in this paper: a continuous and closed-loop dynamic model for simultaneous application of input and output control in WLC. A closed-loop model includes automatic control, so that the job release and the capacity adjustments of the machines are defined in a dynamic way (as function of the shop floor real-time state) and vary in a continuous way.

Three simulations were carried with a control rule that applies input and output control simultaneously (simultaneous release and capacity adjustments). Simulation A is conducted to study the transient response of the system. Simulations B and C include fluctuations in order entry/demand according to a step and a sinusoidal signal.

The implemented feedback control leads the buffer levels to achieve the defined targets and the processing frequencies to stabilize in simulations A and B. In Simulation C, the system responded to the cyclic variations of demand with a maximum 15 % variation of capacity. Even with an increase of almost 20 % in the order entry signal, the WIP did not increase significantly. This means shorter throughput times and a balanced and predictable shop floor, reactive to disturbances. In the proposed closed-loop model, the adjustments are a function of the real-time state of the system.

The first contribution of this paper is the proposition of a prescriptive model that indicates when to release orders and to which amount to adjust the system's capacity. A managerial analysis should define how these adjustments shall be implemented, and its viability. Another relevant contribution is the simultaneous application of input and output control in WLC, considering that this kind of study is scarce in the literature.

In the literature review, no model of WLC that applies the concept of feedback control – as seen in Control Theory – was found. Thus, there is still space for the implementation of continuous and closed-loop simulations of Workload Control, bringing relevant information about the dynamic behaviour of the system. Interdisciplinary models based on the integration of Control Theory and Operations Management can lead to smart production control systems, and expands the range of tools to be used in Production Engineering and Manufacturing research.

For future research, the proposition of different scenarios (changes in the initial/reference values, use of different signal shapes for the order entry) and the development and simulation of different control rules for both input and output control are suggested. This will allow studying the system's behaviour when the controllers have a more global view of the system. This paper is a proof of concept of the closed-loop simulation of WLC. Thus, future studies could apply this model to more complex real systems. Another suggestion is the implementation of feedback loops to models based on discrete event simulation (DES). This is not trivial to be developed.

Acknowledgement

The second author acknowledges the financial support from the São Paulo Research Foundation under process #2019/12023-1 and from the National Council for Scientific and Technological Development under process #310812/2021-6. The third author acknowledges the financial support from the Slovenian Research and Innovation Agency (ARIS) (research core funding No. P2-0219).

References

[1] Shaladdin Muda, M., Hendry, L. (2003). The SHEN model for MTO SMEs: A performance improvement tool, *International Journal of Operations & Production Management*, Vol. 23, No. 5, 470-486, doi: 10.1108/01443570310471820.

[2] Lödding, H. (2013). *Handbook of manufacturing control: Fundamentals, description, configuration*, Springer, Heidelberg, Germany, doi: 10.1007/978-3-642-24458-2.

[3] Breithaupt, J.-W., Land, M., Nyhuis, P. (2002), The workload control concept: Theory and practical extensions of load oriented order release, *Production Planning & Control,* Vol. 13 No. 7, 625-638, doi: 10.1080/0953728021000026230.

[4] Sagawa, J.K., Land, M.J. (2018). Representing workload control of manufacturing systems as a dynamic model, *IFAC-PapersOnLine*, Vol. 51 No. 2, 825-830, doi: 10.1016/j.ifacol.2018.04.016.

[5] Duffie, N.A., Roy, D., Shi, L. (2008). Dynamic modeling of production networks of autonomous work systems with local capacity control, *CIRP Annals,* Vol. 57, No. 1, 463-466, doi: 10.1016/j.cirp.2008.03.018.

[6] Ferney, M. (2000). Modelling and controlling product manufacturing systems using bond-graphs and state equations: Continuous systems and discrete systems which can be represented by continuous models, *Production Planning & Control*, Vol. 11, No. 1, 7-19, doi: 10.1080/095372800232441.

[7] Sagawa, J.K., Nagano, M.S., Speranza Neto, M. (2017). A closed-loop model of a multi-station and multi-product manufacturing system using bond graphs and hybrid controllers, *European Journal of Operational Research*, Vol. 258, No. 2, 677-691, doi: 10.1016/j.ejor.2016.08.056.

[8] Rossit D.A., Tohmé, F., Frutos, M. (2019). Industry 4.0: Smart scheduling, *International Journal of Production Research*, Vol. 57, No. 12, 3802-3813, doi: 10.1080/00207543.2018.1504248.

[9] Cañas, H., Mula, J., Campuzano-Bolarín, F., Poler, R. (2022). A conceptual framework for smart production planning and control in Industry 4.0, *Computers and Industrial Engineering*, Vol. 173, Article No. 108659, doi: 10.1016/j.cie.2022.108659.

[10] Thürer, M., Stevenson, M., Land, M.J., Fredendall, L.D. (2019). On the combined effect of due date setting, order release, and output control: An assessment by simulation, *International Journal of Production Research*, Vol. 57, No. 6, 1741-1755, doi: 10.1080/00207543.2018.1504250.

[11] Thürer, M., Stevenson, M., Land, M.J. (2016). On the integration of input and output control: Workload control order release, *International Journal of Production Economics*, Vol. 174, 43-53, doi: 10.1016/j.ijpe.2016.01.005.

[12] Wu, Z., Weng, M.X. (2005). Dynamic due date setting and shop scheduling for make-to-order companies, In: *Proceedings of the IIE Annual Conference*, Institute of Industrial and Systems Engineers (IISE), USA, 1-6.

[13] Weng, M.X., Wu, Z., Qi, G., Zheng, L. (2008). Multi-agent-based workload control for make-to-order manufacturing, *International Journal of Production Research*, Vol. 46, No. 8, 2197-2213, doi: 10.1080/00207540600969758.

[14] Akillioglu, H., Ferreira, J., Onori, M. (2013). Demand responsive planning: Workload control implementation, *Assembly Automation*, Vol. 33, No. 3, 247-259, doi: 10.1108/AA-12-2013-040.

[15] Kingsman, B.G., Tatsiopoulos, I.P., Hendry, L.C. (1989). A structural methodology for managing manufacturing lead times in make-to-order companies, *European Journal of Operational Research*, Vol. 40, No. 2, 196-209, doi: 10.1016/0377-2217(89)90330-5.

[16] Glavan, M., Gradišar, D., Strmčnik, S., Mušič, G. (2013). Production modelling for holistic production control, *Simulation Modelling Practice and Theory*, Vol. 30, 1-20, doi: 10.1016/j.simpat.2012.07.010.

[17] Andonovski, G., Mušič, G., Blažič, S., Škrjanc, I. (2018). Evolving model identification for process monitoring and prediction of non-linear systems, *Engineering Applications of Artificial Intelligence*, Vol. 68, 214-221, doi: 10.1016/j.engappai.2017.10.020.

[18] Ožbot, M., Lughofer, E., Škrjanc, I. (2023). Evolving neuro-fuzzy systems based design of experiments in process identification, *IEEE transactions on fuzzy systems*, Vol. 31, No. 6, 1995-2005, doi: 10.1109/tfuzz.2022.3216992.

[19] Tomažič, S., Andonovski, G., Škrjanc, I., Logar, V. (2022). Data-driven modelling and optimization of energy consumption in EAF, *Metals*, Vol. 12, No. 5, Article No. 816, doi: 10.3390/met12050816.

[20] Xanthopoulos, A.S., Chnitidis, G., Koulouriotis, D.E. (2019). Reinforcement learning-based adaptive production control of pull manufacturing systems, *Journal of Industrial and Production Engineering*, Vol. 36, No. 5, 313-323, doi: 10.1080/21681015.2019.1647301.

[21] Ragazzini, L., Negri, E., Macchi, M. (2021). A digital twin-based predictive strategy for workload control, IFAC-PapersOnLine, Vol. 54, No. 1, 743-748, doi: 10.1016/j.ifacol.2021.08.183.
[22] Vespoli, S., Mattera, G., Marchesano, M.G., Nele, L., Guizzi, G. (2025). Adaptive manufacturing control with deep reinforcement learning for dynamic WIP management in Industry 4.0, *Computers & Industrial Engineering*, Vol. 202, Article No. 110966, doi: 10.1016/j.cie.2025.110966.
[23] Ikeziri L.M., de Souza, F.B., da Silva Meyer, A., Gupta, M.C. (2021). Pulling the distribution in supply chains: Simulation and analysis of Dynamic Buffer Management approach, *International Journal of Systems Science: Operations & Logistics*; Vol. 10, No. 1, Article No. 1981480, doi: 10.1080/23302674.2021.1981480.
[24] Oosterman, B., Land, M., Gaalman, G. (2000). Influence of shop characteristics on workload control, *International Journal of Production Economics*, Vol. 68, No. 1, 107-119, doi: 10.1016/S0925-5273(99)00141-3.
[25] Sagawa, J.K., Mušič, G. (2019). Towards the use of bond graphs for manufacturing control: Design of controllers, *International Journal of Production Economics*, Vol. 214, 53-72, doi: 10.1016/j.ijpe.2019.03.017.
[26] Sagawa, J.K., Nagano, M.S. (2015). Applying bond graphs for modelling the manufacturing dynamics, *IFAC-PapersOnLine*, Vol. 48, No. 3, 2047-2052, doi: 10.1016/j.ifacol.2015.06.390.
[27] Resman, M., Pipan, M., Šimic, M., Herakovič, N. (2019). A new architecture model for smart manufacturing: A performance analysis and comparison with the RAMI 4.0 reference model, *Advances in Production Engineering & Management*, Vol. 14, No. 2, 153-165, doi: 10.14743/apem2019.2.318.
[28] Broenink, J.F. (1999). 20-SIM software for hierarchical bond-graph/block-diagram models, *Simulation Practice and Theory*, Vol. 7, No. 5-6, 481-492, doi: 10.1016/S0928-4869(99)00018-X.
[29] Lin, J., Naim, M.M., Purvis, L., Gosling, J. (2017). The extension and exploitation of the inventory and order based production control system archetype from 1982 to 2015, *International Journal of Production Economics*, Vol. 194, 135-152, doi: 10.1016/j.ijpe.2016.12.003.

APEM journal

Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 43–60
https://doi.org/10.14743/apem2025.1.526

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Configuring supply chain governance and digital capabilities for resilience: Evidence from the manufacturing sector

Liang, P.P.[a,*], Jili, H.H.[a], Lv, Y.Q.[a,*], Xu, Y.[a]

[a]School of Economics and Management, Anqing Normal University, Anqing, P.R. China

A B S T R A C T

In an increasingly complex and turbulent global environment, achieving resilience in manufacturing supply chains has become a critical strategic priority. Drawing on a sample of 300 manufacturing firms, this study examines both the net and configurational effects of supply chain governance mechanisms and dynamic digital capabilities on supply chain resilience. Using structural equation modeling and fuzzy-set qualitative comparative analysis (fsQCA), the findings reveal that: Contractual governance, relational governance, digital sensing capability, digital resource integration, digital-driven innovation, and digital-enabled business capabilities each have a positive impact on manufacturing supply chain resilience. In the overall sample, only relational governance demonstrates a relatively strong individual effect, while none of the six governance or digital capability dimensions serve as necessary conditions for high resilience in subsample analyses. For high-tech manufacturing firms, two resilient configurations are identified: 1) basic digital enablement with strong governance synergy, and 2) advanced digital enablement with strong governance synergy. In contrast, non-high-tech firms exhibit three distinct resilient configurations: 1) digital integration–driven, 2 advanced digital enablement with relational governance dominance, and 3) dual-core digital enablement with robust governance synergy. These insights provide nuanced theoretical contributions and practical implications for configuring governance and digital strategies to build sustainable supply chain resilience in the manufacturing sector.

A R T I C L E I N F O

Keywords:
Manufacturing supply chain;
Resilience;
Supply chain governance;
Dynamic digital capability;
Fuzzy-set qualitative comparative analysis (fsQCA);
Configuration analysis

**Corresponding author:*
614lpp@163.com
(Liang, P.P.)
lvyouqing2011@163.com
(Lv, Y.Q.)

Article history:
Received 12 November 2024
Revised 16 April 2025
Accepted 20 April 2025

1. Introduction

Manufacturing supply chain resilience (MSCR) is intrinsically linked to national security and economic development. Over the past decade, manufacturing supply chains (MSCs) have been increasingly vulnerable to external shocks, including trade protectionism, technological embargoes, and global public health crises. Unanticipated disruptions—such as core component cutoffs due to sanctions—can severely impair operational performance and even threaten the survival of individual firms, triggering cascading effects across the entire supply chain network. The specialization and fragmentation of production processes have further increased systemic risks, exacerbated by the need for tighter coordination and real-time demand-supply matching.

Numerous real-world cases underscore the disruptive potential of such events. For instance, the 2023 strikes in France—sparked by public dissatisfaction with pension reforms—caused major disruptions to both maritime and inland freight flows. Similarly, the COVID-19 pandemic dramatically reshaped global supply chain structures. According to Dun & Bradstreet, up to 94 per cent of the world's top 1,000 companies experienced supply chain disruptions, with the automotive and electronics manufacturing sectors being the most affected. While some disruptions may be manageable in the short time, others pose significant threats to the long-term resilience and competitiveness of MSCs. As a result, the development of robust, long-term resilience mechanisms has garnered increasing attention from both industry practitioners and academic researchers.

Largely, previous organizations relied significantly on supply chain governance (SCG) to mitigate disruptions and build resilient supply chains. Specific SCG measures can fall into two broad categories, namely contractual governance and relational governance. The former aligned with transaction cost theory emphasizes the use of contracts to safeguard against opportunism and conflict. The latter grounded in the social exchange theory (SET) and relational exchange theory, aiming to curb opportunism by instituting relationship-based norms and developing trust – building mechanisms. Nevertheless, the inherent looseness of supply chain structures, coupled with the bounded rationality of decision-makers, amplifies the vulnerability to opportunistic risks. This, in turn, undermines the efficacy of supply chain governance, particularly within dynamically evolving environments.

According to the dynamic capabilities view (DCV) proposed by Teece *et al*., organizations often develop dynamic capabilities to mitigate the impact of unexpected risks on supply chain performance [1], particularly under highly competitive pressures and in dynamic environments, by integrating, building, and reconfiguring resources. Based on Dubey *et al*. [2] we argue that dynamic capabilities are multi-faceted, encompassing both the ability to capture new opportunities and risks and the ability to utilize available resources and technologies to address them. Crucially, firms are not necessarily strong across all types; the appropriate response to supply chain disruptions is to leverage the specific competencies in which they excel. Moreover, digital-enabled technologies as a key option in crisis scenarios play a significant role in improving supply chain resilience. To maintain competitiveness during turbulent times, organizations are required to develop their digital capabilities for enhancing supply chain resilience to remain competitive in the digital era.

Integrating the above two perspectives, MSCR features multiple concurrent causalities and encompasses different levels. This necessitates a configuration perspective to uncover the multiple equivalent configurations that build supply chain resilience. To better address issues such as “enterprises being ‘willing but unable’ when facing risks” or “possessing strong dynamic capabilities yet remaining "powerless to reverse the situation”, this study incorporates supply chain governance into the configuration analysis framework and matches it with dynamic capability. This approach aims to explore the influencing factors and their configuration mechanisms of supply chain resilience. The main problems to be solved in this paper are as follows.

- How does supply chain governance initiative and dynamic digital capabilities affect MSCR?
- Which factor configurations may constrain MSCR?
- What paths to achieving high MSCR with different technological levels?

Compared with the extant literature, this study makes three primary contributions:

- We develop and empirically validate a theoretical framework elucidating the synergistic mechanisms through which supply chain governance and dynamic capabilities jointly enhance MSCR.
- Distinguishing from net effect studies, we innovatively adopted the fsQCA approach to explore the configuration effect of multiple factors on MSCR, in response to the call from academics for mixed studies of mainstream statistical methods and qualitative comparative analysis methods.
- Considering the "causal complexity" behind supply chain resilience management, the equivalent driving mechanisms for achieving high MSCR (e.g., different routes the same destination) are revealed, which can provide an actionable scenario framework for enhancing MSCR.

The rest of this study is organized as follows. Section 2 presents the theoretical framework and the hypotheses development. Section 3 outlines the research methodology, including the questionnaire design and data-collection process. In Section 4, we conduct an empirical analysis of MSCR using SEM, while Section 5 examines the MSCR mechanism through a hybrid approach combining NCA and fsQCA. Finally, the main findings and conclusions are presented and discussed in the final two Sections.

2. Theoretical framework and research hypotheses

2.1 Supply chain resilience (SCR)

SCR refers to the ability of interconnected supply chain enterprises to maintain their own system stability and avoid chain breakage when exposed to internal and external shock risks, as well as the ability to anticipate and react to future uncertainty [3]. Subsequently, some scholars have further extrapolated this concept across the dimensions: recovery from disruptions, risk resistance, and complexity adaptation [4]. The SCR measurement metrics can be categorized into four groups, respectively: core capability indicators (e.g., supply chain flexibility, visibility agility), recovery metrics (degree/time to restore original state), financial performance, and network topology metrics [5, 6].

Currently, the research paradigm on SCR strategies has evolved from "static to dynamic" and" traditional to complex". Early studies grounded in the static resource-based theory (RBT), emphasized cooperation production, supply chain network structure design, supply chain redundancy design, contract design and governance [7], etc. To address RBT's limitations in analyzing technological shifts, changing consumer preferences, and dynamic competition, scholars have begun to apply DCV to reveal the antecedents, processes and outcomes of SCR [8]. The DCV posits that mere possession of scarce resources is insufficient for competitive advantages—these resources must be reconfigured and deployed effectively.

2.2 SCG and MSCR

Contractual governance relies on written agreements to regulate relationships among manufacturing supply chain members. These contracts include explicit terms that clearly define the responsibilities and obligations of each party [9]. When unexpected disruptions occur, clearly defined responsibilities help prevent task shirking and interpersonal conflicts. They also facilitate effective information sharing through standardized parameters such as price, quantity, logistics, and quality, thereby enhancing the efficiency of uncertainty management [10]. Moreover, comprehensive contracts address a broad spectrum of potential risks and corresponding countermeasures. This provides partners with predetermined rules and procedures, which in turn reduces decision-making uncertainty and promotes supply chain stability. Legally enforceable contracts also ensure compliance, as violations trigger timely corrective actions or penalties. This mechanism deters opportunistic behavior during crises and strengthens the supply chain's systemic resistance to risk [11].

Hypothesis H1a: Contractual governance positively affects MSCR.

Relational governance emphasizes coordinating each other's behaviors and developing long-term relationships, through the construction of social mechanisms such as trust, commitment and reciprocity [12]. As a cornerstone of social exchange, trust motivates supply chain partners to share critical information and resources, facilitating joint actions for rapid operational adaptation [13]. Consequently, institutionalizing trust mechanisms is pivotal for cultivating risk-resilient manufacturing supply chain. Relational commitment as another ingredient of SET, instills confidence in manufacturing supply chain members, put forth the essential effort and enhances MSCR by creating reciprocally beneficial exchanges [14]. According to SET, reciprocity is mutual exchanges that partners consider fair and provide long term gratification because behavior by an exchange partner will encourage reciprocal action by other partners. From a long-term view, reciprocity mechanisms can significantly enhance MSCR by facilitating resource sharing, risk sharing, and collaborative innovation, enabling partners to better cope with uncertainties [15, 16].

Hypothesis H1b: Relational governance positively affects MSCR.

2.3 Dynamic digital capability and MSCR

Digital capability is commonly defined as the abilities endowed by digital technologies that respond quickly to environmental changes. Following Teece *et al*. [17] and Sousa-Zomer *et al*. [18], a defining attribute of digital technologies amid continuous disruption is their proactive environmental scanning capability. Digital sensing capability refers to an organization's ability to collect, analyze and interpret digital information from its internal and external environments [19]. This helps with scanning the external environment for unexpected disruptions and taking preventive actions. For instance, manufacturers with strong digital sensing can predict a sudden surge in demand for a particular product and adjust production accordingly. They can also minimize the impact of disruptions on their supply chain, by proactively managing inventory, adjusting production schedules, and collaborating with suppliers. Hence, digital capabilities must possess the seizing ability.

Hypothesis H2a: Digital sensing capabilities positively affect MSCR.

Sirmon and Hitt [20] argue that resource integration refers to an organization's ability to create economic value by assembling, combining, optimizing and rationally allocating the internal and external resources. Digital resources are a key source for building dynamic digital capabilities. This dimension focuses on the ability to combine and optimize digital resources across the entire manufacturing supply chain, involving integrating data from different systems and processes to create a unified operational view [21]. When manufacturers are capable of effectively integrating digital resources, they can eliminate redundancies, enhance communication among partners, and respond swiftly to changes, resulting in improved coordination and adaptability amid disruptions. Digital resource integration capability focuses on data management, resource orchestration and process integration, whereas digital-driven innovation capability concentrates on how digital technologies can be leveraged to drive innovation [22]. In the manufacturing sector, digital-driven innovation refers to embedding digital technologies into the manufacturing process to drive improvements and create new opportunities [23]. This form of innovation goes beyond simply adopting digital tools, it entails a fundamental transformation in how manufacturing operations are conceived, executed, and managed. For instance, Internet of Things (IoT) sensors collect real-time data from devices and installations, providing valuable insights to make decisions for optimizing processes, predicting maintenance needs and addressing quality deviations. What's more, the innovative application of blockchain technology enables greater transparency and traceability in manufacturing supply chains, thereby reducing the risk of disruption.

Hypothesis H2b: Digital resource integration capacity positively affects MSCR.

Hypothesis H2c: Digitally- driven innovation capability positively affects MSCR.

In practice, numerous manufacturing firms have successfully leveraged digital technology to achieve business transformation, and digital-driven business capability has gained significant attention from organizational scholars [24]. As an important component of dynamic digital capability, digital-driven business capability refers to an organization's proficiency in utilizing digital technologies, data resources and digital mindset to drive business growth, optimization and transformation [25]. This capability manifests in various ways, including innovating business models, formulating more effective marketing strategies, expanding sales channels and customer bases, and optimizing business resourcing through various digital means [26, 27]. For instance, supported by digital technologies, firms can generate new value growth through business transformation, collaboratively address uncertainties and risks by breaking down information silos, and improve operational efficiency by creating a more agile MSCs network. It further helps to enhance MSCR through improved agility and resistance.

Hypothesis H2d: Digitally-driven business capability positively affects MSCR.

The SEM-based theoretical model is shown in Fig. 1. Moreover, it's clear that a dynamic process of developing supply chain capacity can enhance MSCR. However, some cases disclose that not all supply chains with dynamic digital capabilities are necessarily capable of actively tackling supply chain disruptions. Taking Motorola, once a leading player in the mobile phone industry, as an example, exhibited poor coordination with suppliers during the product design and manufacturing phases and was incapable of responding promptly to market demands. Eventually, it underwent multiple acquisitions and restructurings.

This implies that even if an enterprise possesses outstanding dynamic digital capabilities, without choosing effective governance initiatives, it can still lead to the enterprise's cessation of operation. Consequently, scholars have increasingly recognized that the dominant role of SCG initiatives should not be ignored and have attempted to deeply explore the internal mechanisms of MSCR from the relationship management perspective [28]. As a matter of fact, dynamic digital capabilities can provide more advanced tools and means for supply chain governance. Conversely, appropriate supply chain governance initiatives can facilitate the effective application and integration of digital resources, thereby avoiding resource waste and information silos. These two dimensions exhibit a superior synergistic effect, which has not been adequately considered in existing research [2, 29]. Therefore, this paper proposes a conceptual model of MSCR from the configurational perspective, as presented in Fig. 2.

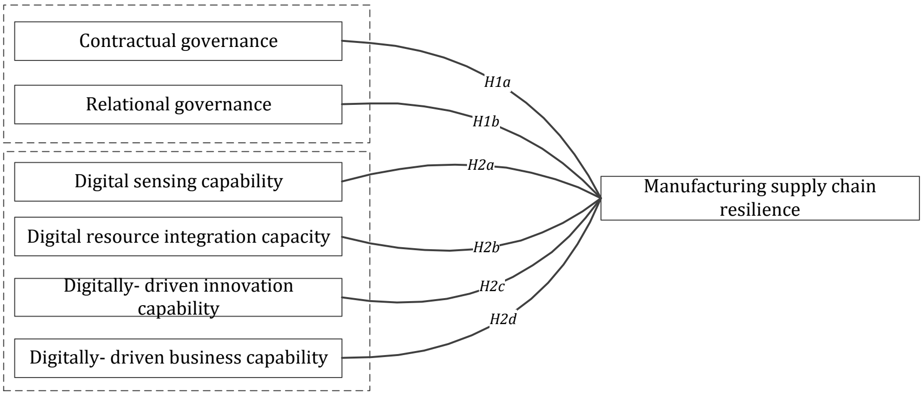


Fig. 1 The SEM-based theoretical model

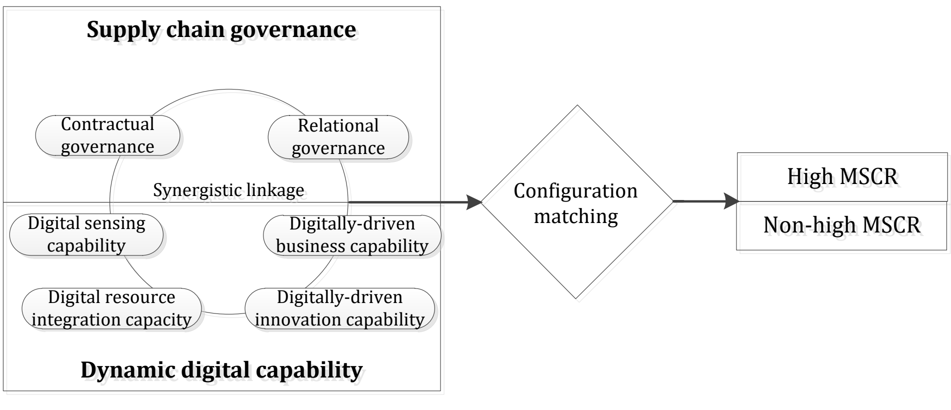


Fig. 2 Conceptual model based on a configurational perspective

3. Research methodology

3.1 Method of hybrid NCA and fsQCA

As a case-oriented method, Qualitative Comparative Analysis (QCA) is designed to capture the causal complexity and interdependence among multiple conditions in configurational research. Among its variants, fsQCA has emerged as a mainstream analytical paradigm, as it accommodates continuous variables and partial membership, thereby enhancing both analytical practicality and generalizability. However, fsQCA primarily identifies configurations of antecedent conditions associated with an outcome, offering qualitative insight rather than quantitative assessment of the degree to which specific conditions are necessary. Especially in fuzzy-set contexts, necessity is not a binary concept ("yes" or "no") but rather a matter of degree. To address this limitation, Necessary Condition Analysis (NCA) can be integrated with fsQCA. NCA quantitatively assesses the extent to which a condition is necessary for a particular outcome, thereby complementing fsQCA and enriching the explanatory power of social science theories. This hybrid approach significantly enhances both descriptive precision and theoretical robustness.

This study begins by using SEM to explore the effects of dynamic digital capabilities and SCG initiatives on MSCR. Following this, the NCA method is employed to identify whether certain dimensions of digital capabilities or SCG initiatives constitute necessary conditions for high MSCR, and if so, to what degree. Concurrently, the QCA approach is applied to verify the robustness of these necessary condition findings. Finally, fsQCA is utilized to delve into the complex causal mechanisms through which dynamic digital capabilities and SCG initiatives shape high levels of MSCR. A heterogeneity analysis is also conducted across industries with differing levels of technological sophistication. As a configurational method, fsQCA conducts cross-case comparative analysis from a holistic perspective. It aims to uncover which combinations of conditions lead to the presence—or absence—of the outcome. This approach is well-suited for examining the multifactorial and complex formation mechanism of MSCR.

3.2 Questionnaire design

We employ the questionnaire survey data from manufacturing firms to study the impact of the configuration mechanism between dynamic digital capabilities and SCG initiatives on MSCR. To ensure the sample reliability and validity, this questionnaire mainly draws on the mature content that has been published in domestic and foreign literature (as shown in Table 1). Primary data collection utilized a five-point Likert scale, where values ranging from 1 ("strongly disagree") to 5 ("strongly agree"), capturing progressive agreement levels across all variables.

Table 1 Measurement items of each variable

Constructs	Measurement items	References
Contractual governance (CG)	CG1: Sign an agreement with supply chain partners. CG2: The agreement improves product quality. CG3: The agreement ensures that both sides understand the product. CG4: The agreement improves communication efficiency with partners.	[30, 31]
Relational governance (RG)	RG1: Have close cooperative relationship with supply chain partners. RG2: Share the long-term and short-term plans with partners. RG3: Trust the commitments made by partners.	[31]
Digital sensing capability (DSC)	DSC1: Accurately predict industry technology trends leveraging digital technology. DSC2: Fully track the changes and trends of customer needs leveraging digital technology. DSC3: Identify opportunities brought by competitive changes leveraging digital technology. DSC4: Identify opportunities brought by supply and demand changes (e.g., changes in supplier quotations, emerging supply markets, and changes in consumer preferences) leveraging digital technology.	[2]
Digital resource integration capability (DRIC)	DRIC1: Be able to effectively achieve the transfer and combination of digital resources. DRIC2: Be able to effectively allocate and utilize data resources. DRIC3: Be able to obtain abundant data resources from the supply chain network.	[22, 32]
Digitally-driven innovation capability (DIC)	DIC1: Have a high tolerance for losses stemming from innovation. DIC2: Use digital means to introduce more new products and services. DIC3: Use digital means to continuously improve the manufacturing process. DIC4: Use digital means to transform production mode at a faster speed.	[33]

Table 1 (Continuation)

Digitally-driven business capability (DBC)	DBC1: Enabled by digital technology, we can promptly execute counter-measures once major competitors target our customers with promotional activities. DBC2: Digital technology empowers us to execute timely and effectively marketing strategies. DBC3: Leveraging digital technology, we can proficiently acquire and assimilate fundamental and pivotal business technologies. DBC4: Digital technology facilitates the continuous development of initiatives aimed at reducing production costs. DBC5: Digital technology allows for the efficient organization of production processes. DBC6: Digital technology enables the efficient allocation of resources across production and other departments.	[25]
Manufacturing supply chain resilience (MSCR)	MSCR1: Preparedness for potential disruption impacts across the supply chain. MSCR2: Rapid respond to supply chain disruption events MSCR3: Maintain basic business operations in the event of disruption. MSCR4: Preserve the desired level of control over structure and function in the event of disruption. MSCR5: Recover speed to its original state after being disrupted. MSCR6: Adaptive transformation to an improved post-disruption state.	[34]

3.3 Sample selection and data collection

To empirically test the proposed hypotheses, data were collected from manufacturing firms in China. The survey participants included top and middle managers, and confidentiality of their responses was strictly maintained. The qu items were adapted from validated measurement scales in prior research, with item wording carefully adjusted to align with our research context. A pilot test was conducted with 20 enterprises to finalize the questionnaire, which was refined for a large-scale distribution. These procedures ensured reliability and validity. Data collection employed both field research and online distribution methods, yielding a total of 300 valid responses, including 102 from field surveys. Sample characteristics are summarized in Table 2, while Table 3 presents the descriptive statistical for all variables.

The classification of industry technology level follows the OECD's high-tech industry classification standard, aligned with China's Industrial Classification of National Economy (GB/T 4754-2017). According to this criterion, manufacturing sectors with relatively high R&D intensity are categorized as high-tech manufacturing industries. These encompass six major categories: aerospace vehicle and equipment manufacturing, electronic and communication equipment manufacturing, computer and office equipment manufacturing, pharmaceutical manufacturing, medical equipment and instrument manufacturing, and information chemical manufacturing. The remaining industries are classified as non-high-tech manufacturing industries.

Table 2 Descriptive statistics of the sample

	Sample characterization	Norm	Sample size	Percentage (%)
Firm information	Firm size (no. of employees)	≤ 50	7	2.3
		51-200	63	21.0
		201-500	114	38.0
		501-1000	74	24.7
		> 1000	42	14.0
	Firm age	1-3	10	3.3
		4-6	33	11.0
		7-9	62	20.7
		≥ 10	195	65.0
	Industry technology level	High-technology	104	34.7
		Non-high-technology	196	65.3
Respondent information	Educational attainment	Below bachelor's degree	18	6
		Bachelor's degree	235	78.3
		Master's degree or above	47	15.7
	Current position	Top manager	20	6.7
		Middle manager	280	93.3

Table 3 Descriptive statistical analysis of variables

Statistical indicators	Antecedent condition						Outcome variable
	Supply chain governance		Dynamic digital capability				Manufacturing supply chain resilience
	CG	RG	DSC	DRIC	DIC	DBC	/
Average value	4.242	4.231	3.970	4.148	3.961	4.176	3.968
Median value	4.500	4.667	4.250	4.667	4.200	4.500	4.333
Standard deviation	0.770	0.874	0.944	0.921	0.806	0.890	0.927
Minimum value	1.250	1.333	1.250	1.333	1.400	1.500	1.500
Maximum value	5.000	5.000	5.000	5.000	5.000	4.833	5.000

4. Empirical analysis of MSCR mechanism based on SEM

4.1 Reliability and validity

The reliability of the scale data is typically assessed using two indicators: internal consistency coefficient (Cronbach's α coefficient) and composite reliability (CR value). In this study, SPSS 26.0 software was used for analysis. The results presented in Table 4 show that the Cronbach's α and CR value for all variables exceed 0.8, indicating that the scale used in this study has good reliability.

Validity analysis includes four aspects: content validity, structural validity, convergent validity, and discriminant validity. Content validity has been addressed previously. Structural validity, convergent validity, and discriminant validity were examined using confirmatory factor analysis (CFA) with Amos 26.0 software for testing. Prior to CFA, KMO and Bartlett's sphericity test were firstly carried out by using SPSS 26.0 software. The results of the KMO and Bartlett's sphericity test indicate that the data sample is suitable for factor analysis (the KMO value is $0.917 > 0.6$; the Bartlett's sphericity test is significant with $p = 0.000 < 0.05$). CFA results demonstrated a good model fit: $x^2/df = 1.128 < 3, RMSEA = 0.038 < 0.05, GFI = 0.914, NFI = 0.922, CFI = 0.990, IFI = 0.990, TLI = 0.989$, all of which are greater than 0.9, indicating that the model overall fit was good, and the scale had excellent structural validity. The factor loadings of each question item were all greater than 0.6, and the combined reliability CR was greater than 0.8, indicating that the aggregation validity of the scale basically met the standard. The square root of AVE for each variable was greater than the correlation coefficient of that variable with the rest of the variables, indicating that the scale used in this study had good discriminant validity.

Table 4 Reliability and validity analysis

	CG	RG	DSC	DRIC	DIC	DBC	MSCR
CG	0.579						
RG	0.371***	0.644					
DSC	0.379***	0.352***	0.673				
DRIC	0.292***	0.394***	0.371***	0.69			
DIC	0.321***	0.368***	0.382***	0.34***	0.555		
DBC	0.295***	0.353***	0.297***	0.346***	0.336***	0.676	
MSCR	0.462***	0.566***	0.485***	0.508***	0.513***	0.533***	0.641
Cronbach's Alpha	0.845	0.844	0.891	0.867	0.861	0.926	0.914
CR	0.846	0.845	0.892	0.869	0.862	0.926	0.914
The square root of AVE	0.761	0.802	0.820	0.831	0.745	0.822	0.801

4.2 Common method bias test

The data samples used in this study are mainly micro-level data obtained through research. However, a single data source has the potential to cause common method bias and thus affect the research results. In view of this, we apply one-way validation factor analysis to test the data for common method bias using MSCR as a latent factor. As can be seen in Table 5, compared to the original fitted model, the model after the one-way validated factor analysis was poorly fitted and did not meet the reference standard; therefore, this study does not suffer from a serious common method bias problem.

Table 5 Common method bias test

Indicator	One-way validated factor analysis model	Original fitted model	Reference standard
x^2/df	7.285	1.128	<3
GFI	0.510	0.914	>0.9
RMSEA	0.145	0.038	<0.08
CFI	0.504	0.990	>0.9
NFI	0.469	0.922	>0.9
IFI	0.506	0.990	>0.9
RMR	0.130	0.038	<0.05

4.3 Hypothesis testing

We used Amos 26.0 to conduct the structural equation model test. Consequently, we got the path analysis diagram shown in Fig. 3, and the specific results are shown in Table 6. As can be seen from Table 6, CG($\beta = 0.134, p = 0.015$), RG($\beta = 0.247, p = 0.000$), DSC($\beta = 0.165, p = 0.002$), DRIC($\beta = 0.143, p = 0.010$), DIC($\beta = 0.185, p = 0.001$), DBC($\beta = 0.242, p = 0.000$) all have a significant positive impact on MSCR. Thus, hypotheses H1a~H2d are all supported.

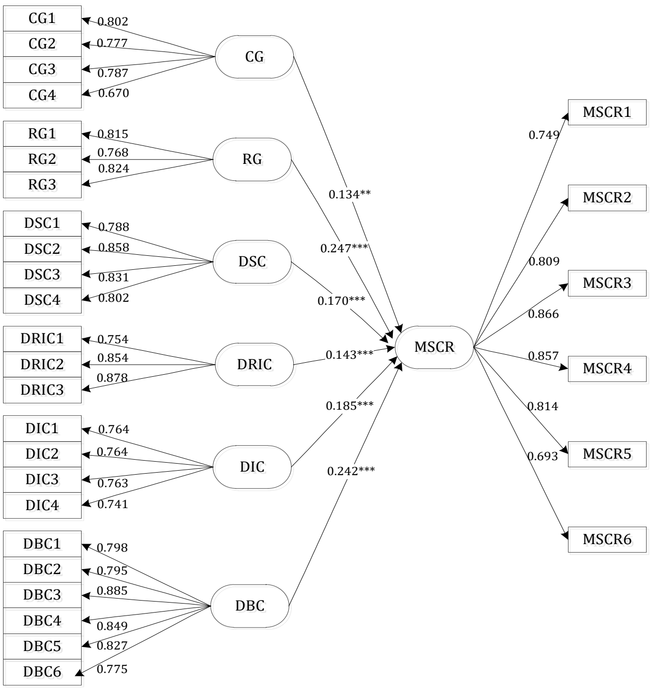


Fig. 3 SEM path diagram

Table 6 Hypothesis testing results

Hypothetical path	Hypothesis	Unstandardized coefficient	Standardized coefficient	S.E.	C.R.	P
CG → MSCR	H1a	0.186	0.134	0.077	2.423	**
RG → MSCR	H1b	0.248	0.247	0.059	4.179	***
DSC → MSCR	H2a	0.165	0.170	0.054	3.089	***
DRIC → MSCR	H2b	0.129	0.143	0.050	2.600	***
DIC → MSCR	H2c	0.201	0.185	0.061	3.302	***
DBC → MSCR	H2d	0.253	0.242	0.056	4.548	***

Notes：** indicates P-value < 0.05；*** indicates P-value < 0.01.

5. MSCR mechanism analysis based on hybrid NCA and fsQCA methods

5.1 Necessary condition analysis

The NCA method adopted in this article can not only identify whether a specific condition is a necessary condition for a certain result but also quantify the effect size of that necessary. The effect size, also termed the bottleneck level, represents the lowest level of a necessary condition required to achieve a particular result, ranging between 0 and 1. Generally, two methods, ceiling regression (CR) and ceiling envelopment (CE), can be used for estimation. A condition is deemed necessary if its effect size (d) is ≥ 0.1 and the permutation-based Monte Carlo simulation test yields a significant result. Table 7 presents the NCA results. Overall, none of the antecedent conditions within dynamic digital capabilities meet both criteria. In terms of SCG, in the total sample, only RG has an effect size greater than 0.1 and a significant result, which is a necessary condition for MSCR with a medium-level effect. For this reason, we conducted a further sub-sample analysis and obtained $d_{CR} = 0.093, d_{CE} = 0.04$ in the high-tech manufacturing sample, and $d_{CR} = 0.038, d_{CE} = 0.046$ in the non-high-tech manufacturing sample. The effect sizes are all less than 0.1, so that CR is not a necessary condition for high MSCR.

Table 8 presents the bottleneck level analysis for antecedent conditions. The results show that there is a bottleneck of dynamic digital capabilities for MSCR level, and to achieve 70 % level (member membership score > 0.7) of MSCR requires 35.1 % level of DBC, 23.5 % level of DIC, and 9.3 % level of RG. No bottleneck effects were found for DRIC, DSC, and CG at this level. However, to reach a high MSCR level of 90 % (membership score > 0.9), higher thresholds are needed: 9.3 % for DRIC, 39.9 % for DBC, 50 % for DSC, 41.2 % for DIC, 9.3 % for RG, and 66.7 % for CG.

In addition, we conducted an analysis of the necessity of individual conditions based on the fsQCA method and further examined the necessary conditions for high MSCR and non-high MSCR in the high-tech and non-high-tech manufacturing industries, and the results are shown in Table 9. The consistencies of the antecedent conditions in the four dimensions of digital dynamic capabilities and the two dimensions of supply chain governance are all less than 0.9, indicating that none of the above eight antecedent conditions are necessary conditions for high MSCR and non-high MSCR.

Table 7 Results of NCA analysis

Antecedent	Estimation method	Ceiling zone	Precision (%)	Effect size	P-value
CG	CR	1.284	96.2	0.098	0.057
	CE	1.708	100	0.130	0.021
RG	CR	1.295	99.0	0.101	0.042
	CE	1.388	100	0.108	0.049
DSC	CR	1.162	99.0	0.095	0.002
	CE	1.535	100	0. 125	0.002
DRIC	CR	1.006	99.0	0.078	0.110
	CE	1.272	100	0.099	0.095
DIC	CR	1.580	98.1	0.133	0.060
	CE	1.932	100	0.162	0.004
DBC	CR	1.539	98.1	0.132	0.095
	CE	2.083	100	0.179	0.000

Notes: (1) The data are the calibrated fuzzy-set membership values. (2) Range of effect size (d): 0 < d < 0.1 is regarded as “small effect”, and 0.1 ≤ d < 0.3 is regarded as “medium effect”. (3) Permutation test (test. rep = 10,000), when p is within the range of less than 0.05, it is significant.

Table 8 Bottleneck level analysis results (%)

MSCR	CG	RG	DSC	DRIC	DIC	DBC
0	NN	NN	NN	NN	NN	NN
10	NN	NN	NN	NN	NN	NN
20	NN	NN	NN	NN	NN	NN
30	NN	NN	NN	NN	NN	NN
40	NN	NN	NN	NN	NN	NN
50	NN	NN	NN	NN	NN	NN
60	NN	NN	NN	NN	NN	NN
70	NN	9.3	NN	NN	23.5	35.1
80	NN	9.3	14.3	9.3	35.3	35.1
90	66.7	9.3	50.0	9.3	41.2	39.9
100	93.3	91.0	92.9	91.0	82.4	95.2

Notes: The analytical method is CR, and NN means "not necessary".

Table 9 Necessary condition test for QCA methodology in high-tech and non-high-tech industries

Antecedent condition		Outcome Variable	
		High MSCR	Non-high MSCR
Supply Chain Governance	CG	0.816 (0.841)	0.599 (0.595)
	~CG	0.417 (0.389)	0.688 (0.672)
	RG	0.714 (0.729)	0.512 (0.553)
	~RG	0.528 (0.562)	0.785 (0.784)
Dynamic digital Capability	DSC	0.778 (0.814)	0.521 (0.583)
	~DSC	0.466 (0.470)	0.780 (0.746)
	DRIC	0.802 (0.711)	0.631 (0.563)
	~DRIC	0.472 (0.623)	0.706 (0.824)
	DIC	0.751 (0.800)	0.478 (0.571)
	~DIC	0.449 (0.450)	0.767 (0.719)
	DBC	0.684 (0.866)	0.535 (0.623)
	~DBC	0.625 (0.463)	0.846 (0.759)

Note: The values in parentheses are the analysis results for non-high-tech manufacturing industries.

5.2 Configuration analysis

Using fsQCA 3.0 software, we analyzed and extracted distinct configurational pathways leading to high MSCR, illustrating the principle of equifinality, where different paths lead to the same destination". A total of 104 valid samples were collected from the high-tech industry. Accordingly, the case frequency threshold was set to 3, the original consistency threshold to 0.8, and the PRI consistency standard to above 0.6. For the non-high-tech industry, the case frequency threshold was set to 4, the original consistency threshold to 0.8, and the PRI consistency standard to above. Core conditions in each grouping were identified by comparing the nested relationship between the intermediate and simple solution via counterfactual analysis: if a grouping appears in both the intermediate solution and the simple solution, it is a core condition, and if it appears only in the intermediate solution, it is an auxiliary condition.

(1) Configurational analysis for high-tech manufacturing industries

Table 10 presents the results following the standard QCA configuration format. We identified two distinct configurations (M1a, M1b, and M2) that consistently generate high MSCR. Notably, each configuration demonstrates a consistency scores exceeding 0.9, signifying their status as sufficient conditions for achieving high MSCR. The overall solution coverage is 0.615, surpassing the 0.5 threshold, which highlights the strong explanatory power of these configurations. To succinctly capture the core attributes and highlight the uniqueness of each configuration, we performed a qualitative analysis of representative cases and took the intensity of SCG and the elementary or advanced dynamic digital capability as the "anchors" for naming the configurations.

Table 10 Sufficiency analysis of condition configuration - High-tech industries

Antecedent condition	High MSCR		
	M1a	M1b	M2
CG	●	●	●
RG	●	●	●
DSC	●	●	⊗
DRIC	●	●	●
DIC		•	⊗
DBC	⊗		●
Consistency	0.937	0.975	0.907
Raw coverage	0.402	0.468	0.237
Unique coverage	0.021	0.105	0.028
Solution consistency	0.953		
Solution coverage	0.615		

Notes: ⊗ and ● respectively indicate that the level of antecedent conditions is not high and relatively high; The large circle represents the core condition, and the small circle represents the auxiliary condition; a blank space indicates that the condition is not important for the generation of results.

a) Configuration M1 *(Elementary digital application—Strong governance synergy type).* Configuration M1 demonstrates that core conditions—CG, RG, DSC and DRIC—jointly drive high MSCR with DBC or DIC serving as auxiliary conditions. In this configuration, the attributes of the high-tech industry highlight the critical role of strong governance. It suggests that focal firms in the manufacturing supply chain should foster both CG and RG to leverage complementary governance advantages. Essentially, digital sensing is the initial detection of digital value, while integration is the optimal combination of value carriers (digital resources). Together, these two capabilities establish a foundational level for extracting digital value, paving the way for more advanced capabilities, such as DIC and DBC.

Configuration M1 comprises two distinct paths:

- Path M1a: The antecedent construct is represented as "CG*RG*DSC*DRIC*~DBC".
- Path M1b: The antecedent construct is represented as "CG*RG*DSC*DRIC*DIC".

Configuration M1a exhibits a consistency score of 0.937, with an original coverage of 0.402 and a unique coverage of 0.021. This configuration accounts for approximately 40.2 % of the cases, primarily in highly regulated sectors such as shipbuilding, aviation, aerospace, and equipment manufacturing. In these sectors, a robust governance structure empowers supply chain partners to navigate complex regulatory and market environments, ensuring compliance and transparency throughout the digital transformation process. In this context, DSC plays a vital role. It enables enterprises to swiftly capture shifts in market demand while providing proactive data support for addressing potential risk events. Additionally, DRIC serves as a collaborative tool for enhancing SCR. By facilitating digital resource integration, it fosters synergistic effects that improve the overall elasticity and stability of the manufacturing supply chain.

Configuration M1b features a consistency of 0.975, an original coverage of 0.468, and a unique coverage of 0.105, explaining the largest number of cases at 46.8 %. These cases are mainly concentrated in industries such as electronic and communication equipment manufacturing, computer and office equipment manufacturing, pharmaceutical manufacturing, and medical instrument and device manufacturing. A notable difference between M1a and M1b is that, M1b incorporates DIC as a auxiliary condition, while M1a includes DBC. This difference can be attributed to significant variations in market environment, product characteristics, and innovation demands between these two industry types. The industries covered by configuration M1b are characterized by rapid technological iteration, intense market competition, and a strong orientation toward mass-market consumer products. In this setting, firms must proactively drive innovation and application of digital technologies to sustain their competitive edge. In contrast, sectors such as shipbuilding and aerospace typically produce highly customized products with stable user demands and longer R&D cycles. Consequently, these industries place less direct reliance on DIC and prioritize developing foundational digital capabilities to ensure product reliability and compliance.

b) Configuration M2 *(Dual-core digital-driven—Strong governance synergy type).* The antecedent construct is "CG*RG*DRIC*DBC*~DSC*~DIC". This configuration indicates that CG, RG, DRIC and DBC are core conditions, while DSC and DIC are absent auxiliary conditions in achieving high MSCR. This configuration yields a consistency of 0.907, with raw coverage of 0.237 and unique coverage of 0.028, explaining approximately 23.7 % of the observed cases. These cases are predominantly situated in the information chemical manufacturing industry, which features long, complex supply chains and high sensitivity to fossil fuel price fluctuations. This configuration highlights the importance of digital optimization and execution across production and operational processes. Specifically, the combined strength of DRIC and DBC facilitates agile resource allocation and rapid response under external shocks, such as war, geopolitical conflict, or abrupt price volatility. These digital capabilities enable supply chain nodes to maintain visibility, redistribute constrained resources, and reconfigure operations in real time. When such capabilities are embedded in a governance structure with CG and RG, firms are better positioned to absorb shocks, contain their propagation, and swiftly restore operational continuity. Therefore, the synergy between strong governance redundancy and dual-core digital capability serves as a critical resilience mechanism against severe external disruptions.

(2) Configurational analysis for non-high-tech manufacturing industries

As shown in Table 11, there are three types of configurations (L1a, L1b, L2, and L3) that contribute to high MSCR. All of their consistencies exceed 0.9, indicating that these configurations are sufficient conditions for achieving high MSCR. Additionally, the solution coverage is 0.697, which is significantly above the threshold of 0.5, demonstrating strong explanatory power. Combined with theory and industry cases analysis, we took DRIC, RG and the elementary or advanced dynamic digital capability as the "anchors" for configuration naming that takes into account both integrity and uniqueness.

Table 11 Sufficiency analysis of condition configuration – non-high-tech industries

Antecedent condition	High MSCR			
	L1a	L1b	L2	L3
CG	•	•	•	
RG		•	●	●
DSC		•		⊗
DRIC	●	●		●
DIC	•		●	⊗
DBC	•		●	●
Consistency	0.959	0.950	0.937	0.910
Raw coverage	0.534	0.511	0.520	0.180
Unique coverage	0.070	0.056	0.056	0.018
Solution consistency	0.917			
Solution coverage	0.697			

a) Configuration L1 *(Digital resource integration dominant type*). This configuration highlights the core role of digital resource integration capability in enhancing MSCR, as it effectively breaks down information silos, facilitates collaboration among stakeholders, and optimizes resources. This, in turn, enhances the decision-making flexibility and market responsiveness of the supply chain, thereby improving the resilience of member enterprises in dynamic market environments.

Configuration L1 comprises two distinct paths:

– Path L1a: The antecedent construct is represented as "CG*DRIC*DBC*DIC".
– Path L1b: The antecedent construct is represented as "CG*RG*DSC*DRIC".

Configuration L1a identifies DRIC as the core condition, with DBC, DIC, and CG serving as auxiliary conditions contributing to high MSCR. The consistency of this configuration is 0.959, with a raw coverage of 0.534 and unique coverage of 0.07. This configuration accounts for approximately 53.4 % of the cases, primarily within the agricultural and food processing sectors. These industries operate in a highly demand-driven market, where seasonal variations and fluctuations in consumer preferences directly influence production and inventory decisions. Moreover, growing societal concerns over food safety highlight the pivotal role of DRIC in enabling real-time monitoring and traceability across the supply chain. DBC and DIC function as auxiliary conditions that support enterprises in optimizing processes and driving innovation. However, in such a responsive market, their effectiveness relies on the foundational support provided by DRIC. Additionally, CG ensures collaboration and compliance among supply chain partners, further enhancing MSCR.

Configuration L1b also identifies DRIC as the core condition, but includes DSC, RG and CG serving as auxiliary conditions. This configuration has a consistency of 0.95, with a raw coverage of 0.511 and unique coverage of 0.056. It accounts for approximately 51.1 % of the cases, primarily within the chemical fiber, rubber and plastic manufacturing, non-ferrous metal smelting, and metal manufacturing sectors. These industries share common characteristics, including complex production processes, high dependence on raw materials, and frequent fluctuations in market demand. Therefore, in practical operational management, on one hand, the focus is on leveraging DRIC and DSC to optimize process flows and enhance market responsiveness, aiming to achieve cost reduction, efficiency improvement, and risk mitigation. On the other hand, the collaborative advancement of CG and RG provides a more comprehensive management framework, enhancing the cooperation efficiency and adaptability of supply chain members.

b) Configuration L2 *(Advanced digital-driven—Relationship-oriented governance synergy type).* The antecedent construct is represented as "CG*RG*DIC*DBC", where CG serves as an auxiliary condition, while the others are considered core conditions. The consistency of this configuration is 0.937, with a raw coverage of 0.52 and unique coverage of 0.056. This configuration accounts for approximately 52 % of the cases, primarily within general and specialized equipment, automotive manufacturing, and electrical machinery and equipment manufacturing sectors. In such an industrial environment with complex products and a highly dependent supply chain, DIC and DBC can effectively work in coordination, drive business process reengineering, and facilitate innovation and optimization in product design, production, and services. This is of vital importance for maintaining a competitive advantage and achieving a high MSCR. Additionally, compared with CG, RG can accelerate knowledge flow and promote collaborative innovation, while CG is more of a support for this relationship. This distinction is particularly significant for understanding the condition configuration in which they jointly achieve high resilience with DIC.

c) Configuration L3 *(Dual-core digital-driven—Relationship-prioritized governance synergy type)*, has its antecedent construct represented as "CG*~DSC*DRIC*~DIC*DBC". The consistency of this configuration is 0.91, with a raw coverage of 0.18 and unique coverage of 0.018. Approximately 18 % of the cases can be explained by this configuration, primarily those in the papermaking, paper products and printing industries, as well as in the manufacturing sectors of cultural, educational, sports and entertainment products. In the digital era, these industries regard the assetization of digital resources as both a foundation and a strategic direction for development. At the same time, real-world cases of digital transformation further underscore the necessity of leveraging digital technologies to optimize operations, enhance customer experience, and improve marketing strategies. Thus, it is evident that DRIC and DBC emerge as the dual-core digital drivers for achieving high MSCR. This configuration also suggests that an RG model should be prioritized to enhance the adaptability and flexibility of manufacturing supply chain by building long-term trust and reciprocity among partners, rather than a CG model that focuses only on short-term compliance.

5.3 Test of robustness

QCA is a set-theoretic approach that is considered robust when slight adjustments to the operation, with subset relationships between the results produced, do not change the substantive explanation of the research findings. We evaluated the robustness of the antecedent configuration that achieves high MSCR by increasing the case frequency threshold and consistency. First, the case frequency thresholds for high-tech manufacturing and non-high-tech manufacturing were adjusted upward by 1, resulting in new configurations that are fundamentally subsets of the original configurations, with no significant changes in core conditions. Second, by increasing the consistency from 0.80 to 0.90, the resulting configurations remained consistent with the original configurations, with no changes in consistency or coverage. The robustness test indicates that the results are robust.

6. Discussion

6.1 Interpretation of findings

This study empirically confirms that supply chain governance and dynamic digital capabilities serve as dual drivers of high MSCR. SEM results show both positively contribute to resilience outcomes, while NCA indicates that no single factor—whether contractual governance, relational trust, or digital innovation— is indispensable, highlighting the causal complexity involved. FsQCA further identifies five distinct high-resilience configurations across high-tech and non-high-tech sectors. These results reveal that resilience does not stem from isolated excellence, but emerges from context-specific combinations of governance and capability elements. High-tech firms tend to benefit from strong governance paired with either foundational or advanced digital enablement, whereas non-high-tech firms rely more on digital resource integration and relational governance. Overall, the findings underscore that high MSCR is shaped by configuration fit and strategic alignment, not from uniform solutions. Firms must tailor strategies to their technological and organizational contexts, embracing configuration logic in place of one-size-fits-all models.

6.2 Theoretical implications

Several theoretical implications are noteworthy. First, in contrast to prior empirical summaries and conceptual models, our study verifies the positive influence mechanism of SCG and dynamic digital capability on enhancing MSCR through SEM. It highlights the critical importance of an organization's dynamic digital capabilities in addressing supply chain risks [2], and also emphasizes the strategic value of effective SCG in securing competitive advantage in volatile environments. These findings enrich our understanding of the logical linkages among the multiple dimensions of SCG, dynamic digital capability, and MSCR.

Second, prior research on SCR has predominantly emphasized the net effects of individual factors, often overlooking how multiple resources and capabilities may interact in a configurational manner to drive resilience. Beyond the conventional SCG initiatives, developing dynamic digital capabilities has emerged as a critical option for building SCR in uncertain and turbulent environments. However, existing studies have seldom investigated the synergistic configuration of SCG and digital capabilities within a complex systems framework. By addressing this gap, the present study offers new theoretical insights that enrich and refine the conceptual foundations of SCR, particularly under conditions shaped by digital transformation.

Finally, this research presents the "causal complexity" of constructing MSCR by identifying multiple, equally effective configurations that lead to high resilience. This finding aligns with Fiss's configuration theory, which posits that similar outcomes can emerge from divergent causal paths [34]. It underscores the industry-specific and context-dependent nature of resilience-building strategies. Firms should adopt a configuration mindset to identify and leverage their core capability combinations to address specific market environments and challenges.

6.3 Managerial implications

This research presents three key managerial implications. First, adapting governance models to industry characteristics is essential. Enterprises are advised to select appropriate governance models based on their technological attributes. The industry heterogeneity analysis reveals no universally applicable conditional configuration, indicating that high-tech and non-high-tech manufacturing sectors follow distinct paths toward achieving high MSCR. Specifically, high-tech industries ought to accentuate the synergy between CG and RG to ensure that all stakeholders remain coordinated amidst technological shifts and market fluctuations. Conversely, non-high-tech manufacturing industries should leverage governance methods suited to their resource profiles and capability structures. Actively fostering trust, commitment, and reciprocity through informal governance can further facilitate the efficient flow of knowledge and resources, thereby enhancing the stability and synergy of the supply chain.

Second, firms should prioritize the development of digital capabilities based on strategic needs. In high-tech sectors, investment should focus on DSC and digital DRIC. DSC enables agile detection of technological trends and shifting market demands, while DRIC enhances the coordination of internal and external data to support supply chain synergy. For non-high-tech manufacturing industry, the strategic priority lies in customer value creation. Here, the emphasis should be placed on developing DBC and DRC. By introducing IoT, big data analysis, and other digital technologies to optimize product design, production processes, and service models, firms can foster business model innovation and improve the responsiveness and flexibility of the supply chain.

Third, managers should adopt a configuration-oriented mindset in decision-making. Rather than relying on single-factor approaches, firms should consider how different combinations of SCG and digital capabilities contribute to MSCR. This study reveals that, equivalent configurations may feature either different core conditions or identical core conditions with varying auxiliary conditions. The existence of such multiple-condition configuration reflects the complexity of MSCR management. Accordingly, firms should avoid blindly replicating successful strategies from other contexts. Instead, they should assess their own technological orientation, resource base, and market conditions to develop adaptive, context-specific capability-governance portfolios.

6.4 Limitations and future research

While this study provides valuable theoretical and practical insights, several limitations should be acknowledged. First, the cross-sectional design limits causal inference, limiting the ability to capture temporal dynamics in resilience development. Future research could adopt longitudinal designs or temporal QCA to explore how configurations evolve over time. Second, resource endowment, particularly the constraints faced by small enterprises, represents a critical contextual factor influencing the feasibility of resilience configurations. Due to the limited representation of small firms in the current sample, conducting a dedicated fsQCA for this subgroup was not feasible. Future research should consider expanding the sample size of small firms and incorporate firm size as a moderating variable to further clarify how resource constraints shape configuration selection and performance. Third, cultural and institutional contexts may influence the effectiveness of relational governance mechanisms such as trust and reciprocity, suggesting the need for cross-cultural comparative studies. Finally, the study focuses on governance and digital capabilities, leaving other potential factors—such as policy support or organizational learning—for future exploration. Addressing these areas may enhance the robustness and generalizability of configuration-based resilience research.

7. Conclusion

This study rigorously examines the dual influence of SCG and dynamic digital capabilities on MSCR by integrating SEM, NCA, and fsQCA. Empirical results confirm that both governance mechanisms and digital capabilities significantly enhance MSCR, however, it also reveals inherent causal complexity: no single factor can ensure resilience alone. Five distinct, context-specific configurations across high-tech and non-high-tech sectors were identified, indicating that resilience emerges from tailored combinations of governance and capability elements rather than uniform solutions. These findings deepen theoretical understanding by framing MSCR as a configurational outcome shaped by the interplay of multiple factors under varying contextual conditions. Managerially, they highlight the imperative for firms to align governance models and digital capability development with their specific industry characteristics and strategic priorities, enabling the creation of context-sensitive strategies to withstand supply chain disruptions. Ultimately, this research moves the resilience discourse forward by demonstrating that strategic alignment and configuration fit—not generic prescriptions—are fundamental to sustaining competitive advantage amid escalating supply chain uncertainties.

Acknowledgement

This work is partly supported by Philosophy and Social Science Planning Project of Anhui Province (AHSKQ2020D22), and Scientific Research Foundation of Education Department of Anhui Province (2023AH040066, 2024AH052759).

References

[1] Teece, D.J., Pisano, G., Shuen, A. (1997). Dynamic capabilities and strategic management, *Strategic Management Journal*, Vol. 18, No. 7, 509-533, doi: 10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z.

[2] Dubey, R., Bryde, D.J., Dwivedi, Y.K., Graham, G., Foropon, C., Papadopoulos, T. (2023). Dynamic digital capabilities and supply chain resilience: The role of government effectiveness, *International Journal of Production Economics*, Vol. 258, Article No. 108790, doi: 10.1016/j.ijpe.2023.108790.

[3] Ponomarov, S.Y., Holcomb, M.C. (2009). Understanding the concept of supply chain resilience, *The International Journal of Logistics Management*, Vol. 20, No. 1, 124-143, doi: 10.1108/09574090910954873.

[4] Ivanov, D., Dolgui, A. (2020). Viability of intertwined supply networks: Extending the supply chain resilience angles towards survivability. A position paper motivated by COVID-19 outbreak, *International Journal of Production Research*, Vol. 58, No. 10, 2904-2915, doi: 10.1080/00207543.2020.1750727.

[5] Behzadi, G., O'Sullivan, M.J., Olsen, T.L. (2020). On metrics for supply chain resilience, *European Journal of Operational Research*, Vol. 287, No. 1, 145-158, doi: 10.1016/j.ejor.2020.04.040.

[6] Liang, P.P., Li, C.W. (2019). Impact of cooperation uncertainty on the robustness of manufacturing service system, *Advances in Production Engineering & Management*, Vol. 14, No. 2, 189-200, doi: 10.14743/apem2019.2.321.

[7] Maududy, R., Nurdin, A.M. (2024). An architecture framework for supply chain management systems integrated with supervisory control and data acquisition functionality, *Journal of Logistics, Informatics and Service Science*, Vol. 11, No. 5, 38-51, doi: 10.33168/JLISS.2024.0503.

[8] Kähkönen, A.K., Evangelista, P., Hallikas, J., Immonen, M., Lintukangas, K. (2023). COVID-19 as a trigger for dynamic capability development and supply chain resilience improvement, *International Journal of Production Research*, Vol. 61, No. 8, 2696-2715, doi: 10.1080/00207543.2021.2009588.

[9] Um, K.-H. (2024). Strategic governance dynamics in manufacturing firms: Navigating operational performance through contractual and relational mechanisms in the face of product complexity, *Journal of Manufacturing Technology Management*, Vol. 35, No. 3, 502-523, doi: 10.1108/JMTM-09-2023-0411.

[10] Samant, S., Kim, J. (2023). Best foot forward? The importance of contractual governance mechanisms for innovation from alliances, *Technovation*, Vol. 127, Article No. 102828, doi: 10.1016/j.technovation.2023.102828.

[11] Zhang, Q., Jin, J.L., Yang, D.F. (2020). How to enhance supplier performance in China: Interplay of contracts, relational governance and legal development, *International Journal of Operations & Production Management*, Vol. 40, No. 6, 777-808, doi: 10.1108/IJOPM-02-2020-0093.

[12] Keller, J., Burkhardt, P., Lasch, R. (2021). Informal governance in the digital transformation, *International Journal of Operations and Production Management*, Vol. 41, No. 7, 1060-1084, doi: 10.1108/IJOPM-09-2020-0660.

[13] Giannoccaro, I., Iftikhar, A. (2022). Mitigating ripple effect in supply networks: The effect of trust and topology on resilience, *International Journal of Production Research*, Vol. 60, No. 4, 1178-1195, doi: 10.1080/00207543.2020.1853844.

[14] Yang, Q., Geng, R., Jiang, Y., Feng, T. (2021). Governance mechanisms and green customer integration in China: The joint effect of power and environmental uncertainty, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 149, Article No. 102307, doi: 10.1016/j.tre.2021.102307.

[15] Yang, J., Liu, Y., Kholaif, M.M.N.H.K. (2024). The impact of relationship management on manufacturer resilience in emergencies, *Kybernetes*, Vol. 53, No. 3, 960-989, doi: 10.1108/K-08-2022-1198.

[16] Valiušis, O. (2025). Digitalization modeling of production processes in paper packaging sector companies, *Journal of Management Changes in the Digital Era*, Vol. 2, No. 1, 67-76, doi: 10.33168/JMCDE.2025.0105.

[17] Teece, D., Peteraf, M., Leih, S. (2016). Dynamic capabilities and organizational agility: Risk, uncertainty, and strategy in the innovation economy, *California Management Review*, Vol. 58, No. 4, 13-35, doi: 10.1525/cmr.2016.58.4.13.

[18] Sousa-Zomer, T.T., Neely, A., Martinez, V. (2020). Digital transforming capability and performance: A microfoundational perspective, *International Journal of Operations and Production Management*, Vol. 40, No. 7-8, 1095-1128, doi: 10.1108/IJOPM-06-2019-0444.

[19] Christofi, M., Khan, H., Zahoor, N., Hadjielias, E., Tarba, S. (2024). Digital transformation of SMEs: The role of entrepreneurial persistence and market sensing dynamic capability, *IEEE Transactions on Engineering Management*, Vol. 71, 13598-13615, doi: 10.1109/TEM.2022.3230248.

[20] Sirmon, D.G., Hitt, M.A. (2003). Managing resources: Linking unique resources, management, and wealth creation in family firms, *Entrepreneurship Theory and Practice*, Vol. 27, No. 4, 339-358, doi: 10.1111/1540-8520.t01-1-00013.

[21] Ghosh, S., Hughes, M., Hodgkinson, I., Hughes, P. (2022). Digital transformation of industrial businesses: A dynamic capability approach, *Technovation*, Vol. 113, Article No. 102414, doi: 10.1016/j.technovation.2021.102414.

[22] Zhang, Z.G., Ye, B., Hu, A.T., Chen, L. (2024). How do manufacturing enterprises integrate data resources to enable product innovation performance? The role of calculative and relational inter-organizational trust, *Studies in Science of Science*, Vol. 42, No. 4, 649-659, doi: 10.16192/j.cnki.1003-2053.20230922.003.

[23] Liu, Y., Dong, J.Y., Mei, L., Shen, R. (2023). Digital innovation and performance of manufacturing firms: An affordance perspective, *Technovation*, Vol. 119, Article No. 102458, doi: 10.1016/j.technovation.2022.102458.

[24] Eller, R., Alford, P., Kallmünzer, A., Peters, M. (2020). Antecedents, consequences, and challenges of small and medium-sized enterprise digitalization, *Journal of Business Research*, Vol. 112, 119-127, doi: 10.1016/j.jbusres.2020.03.004.

[25] Lin, J., Lin, S., Benitez, J., Luo, X.R., Ajamieh, A. (2023). How to build supply chain resilience: The role of fit mechanisms between digitally-driven business capability and supply chain governance, *Information & Management*, Vol. 60, No. 2, Article No. 103747, doi: 10.1016/j.im.2022.103747.

[26] Bag, S., Dhamija, P., Luthra, S., Huisingh, D. (2023). How big data analytics can help manufacturing companies strengthen supply chain resilience in the context of the COVID-19 pandemic, *The International Journal of Logistics Management*, Vol. 34, No. 4, 1141-1164, doi: 10.1108/IJLM-02-2021-0095.

[27] Xu, Q., Tan, Z., Zhang, Y. (2024). An evolutionary game analysis of digital decision making in manufacturing enterprises under reward and punishment mechanism, *Economic Computation and Economic Cybernetics Studies and Research*, Vol. 58, No. 1, 52-69, doi: 10.24818/18423264/58.1.24.04.

[28] Faruquee, M., Paulraj, A., Irawan, C.A. (2024). The dual effect of environmental dynamism on proactive resilience: Can governance mechanisms negate the dark side?, *Production Planning & Control*, Vol. 35, No. 15, 2113-2130, doi: 10.1080/09537287.2023.2291378.
[29] Li, Y., Wang, X., Gong, T., Wang, H. (2023). Breaking out of the pandemic: How can firms match internal competence with external resources to shape operational resilience?, *Journal of Operations Management*, Vol. 69, No. 3, 384-403, doi: 10.1002/joom.1176.
[30] Zhou, X., Zhu, Q., Xu, Z. (2023). The role of contractual and relational governance for the success of digital traceability: Evidence from Chinese food producers, *International Journal of Production Economics*, Vol. 255, Article No. 108659, doi: 10.1016/j.ijpe.2022.108659.
[31] Poppo, L., Zenger, T. (2002). Do formal contracts and relational governance function as substitutes or complements?, *Strategic Management Journal*, Vol. 23, No. 8, 707-725, doi: 10.1002/smj.249.
[32] Li, M., Li, Z., Huang, X.D., Qu, T. (2021). Blockchain-based digital twin sharing platform for reconfigurable socialized manufacturing resource integration, *International Journal of Production Economics*, Vol. 240, Article No. 108223, doi: 10.1016/j.ijpe.2021.108223.
[33] Liao, M.C., Jiang, Y.S., Jing, J.M. (2023). Digital innovation capability under innovation ecosystem: Reconceptualization and scale development, *Soft Science*, Vol. 37, No. 5, 62-70, doi: 10.13956/j.ss.1001-8409.2023.05.09.
[34] Aslam, H., Khan, A.Q., Rashid, K., Rehman, S.-U. (2020). Achieving supply chain resilience: The role of supply chain ambidexterity and supply chain agility, *Journal of Manufacturing Technology Management*, Vol. 31, No. 6, 1185-1204, doi: 10.1108/JMTM-07-2019-0263.
[35] Fiss, P.C. (2011). Building better causal theories: A fuzzy set approach to typologies in organization research, *Academy of Management Journal*, Vol. 54, No. 2, 393-420, doi: 10.5465/amj.2011.60263120.

APEM journal

Advances in Production Engineering & Management **ISSN 1854-6250**
Volume 20 | Number 1 | March 2025 | pp 61–74 Journal home: apem-journal.org
https://doi.org/10.14743/apem2025.1.527 Original scientific pape

Independent vs. collaborative blockchain Research and Development: Operational decisions in food supply chains

Kang, Y.X.[a,b]**, Feng, J.Y.**[a,c,*]**, Zhang, W.J.**[a]

[a]School of Economics and Management, Beijing Jiaotong University, Beijing, P.R. China
[b]China Center for Information Industry Development, Beijing, P.R. China
[c]Ministry of Industry and Information Technology, Equipment Industry Development Center, Beijing, P.R. China

A B S T R A C T

As blockchain increasingly demonstrates advantages in enhancing consumer trust and reducing collaborative production costs in food supply chains, more food companies are developing blockchain-based supply chain management platforms. However, in a complex and competitive supply chain environment, food companies face critical operational management challenges in selecting appropriate blockchain R&D approaches and determining optimal R&D levels. This study examines two competing food companies deciding on their optimal food production and blockchain R&D levels. We first established a benchmark model without blockchain adoption. Then, we constructed supply chain operation models for both independent and collaborative blockchain R&D scenarios. By comparing equilibrium decisions across different models, we derived the optimal blockchain R&D model and operational strategies in food supply chains. Furthermore, we extended our analysis to consider asymmetric food substitution scenarios. Our findings revealed that independent blockchain R&D tends to increase equilibrium food production and is more suitable for premium food supply chains. Conversely, collaborative blockchain R&D significantly enhanced overall supply chain profitability. As spillover effects increase, food companies are likely to favor independent blockchain R&D to achieve higher R&D levels and stronger market competitiveness. Additionally, we demonstrated that blockchain R&D levels are influenced by food substitutability and quality credibility.

A R T I C L E I N F O

Keywords:
Blockchain adoption;
Food supply chains;
Game theory;
Competitive collaboration;
Operation strategy;
Spillover effects

**Corresponding author:*
fengjiayi@eidc.org.cn
(Feng, J.Y.)

Article history:
Received 1 October 2024
Revised 31 March 2025
Accepted 7 April 2025

1. Introduction

With the increasing popularity and advancement of blockchain technology, stakeholders in the food supply chain have begun to recognize its value. Food manufacturers, retailers, and blockchain firms are actively engaging in blockchain R&D aimed at enhancing the management and oversight of information across various stages of food production, processing, storage, transportation, and distribution. The primary objective of these initiatives is to establish a decentralized, tamper-proof, and transparent system that improves the integrity and efficiency of the food supply chain.

The benefits of blockchain for the food supply chain are increasingly apparent. For consumers, the adoption of blockchain by food companies enhances engagement by showcasing robust food safety systems, improving brand perception, and fostering confidence in product quality. This increased transparency can stimulate consumer demand by allowing individuals to make more informed purchasing decisions. For food companies, a blockchain network facilitates collaboration between upstream and downstream stakeholders within the supply chain. A distributed ledger maintains comprehensive information on food raw materials, production, processing, storage, logistics, supplier transfers, and retail operations, thereby enabling mutual oversight. Enhanced information sharing across different stages allows companies to quickly assess overall supply chain performance, improve operational efficiency, and make informed decisions while minimizing communication costs. For example, “Blockchain + Vegetables” innovation project in Weifang City, where the Yukesong Digital Agriculture Industry Park developed a digital agricultural management platform that integrates vegetable planting, logistics, storage, processing, and sales. This initiative not only enhanced the quality of agricultural products within the park but also achieved over a 10 % reduction in water and fertilizer usage and a labor cost reduction exceeding 15 %.

Additionally, blockchain R&D has certain positive externalities that can bring multiple benefits to the blockchain industry and society. Firstly, blockchain R&D can promote innovation and advancement in blockchain technology, improving its functionality, performance, and security, thus bringing more value and benefits to the blockchain industry and society. Secondly, it can drive the standardization and normalization of blockchain technology, reducing R&D and operational costs and risks, enhancing credibility and interoperability, and thereby bringing more order and vitality to the blockchain ecosystem and food market. Thirdly, it can facilitate the integration and expansion of blockchain technology with other emerging information technologies such as 5G, IoT, and AI, creating an innovative "Blockchain +" model that supports the food industry in exploring new business operation models and management solutions.

Given the numerous advantages of blockchain, many food companies have begun to invest in blockchain R&D. Common approaches include: (1) Independent R&D, where some food companies choose to develop or customize blockchain solutions to meet their specific business needs. The advantage of independent R&D is that it allows for tailored blockchain solutions based on distinct requirements and scenarios, thereby improving system flexibility and adaptability. However, it requires significant investments in human, material, and financial resources and presents challenges related to immature technology and a lack of standards. (2) Collaborative R&D, where some food companies choose to develop and utilize blockchain solutions in partnership with other organizations to facilitate cross-organization data sharing and collaboration. For instance, global food retailers such as Walmart and Nestlé have partnered with IBM to develop a blockchain-based service system, establishing a global food safety alliance aimed at enhancing the efficiency and transparency of the food supply chain. The advantage of collaborative R&D lies in leveraging the technology and resources of partners, which reduces development and operational costs and risks while increasing system credibility and interoperability. However, it necessitates the coordination of interests and the resolution of issues related to data privacy and security.

Moreover, the R&D costs represent a major obstacle for food companies in adopting blockchain services. For example, in terms of software, blockchain adoption requires selecting or developing an appropriate blockchain platform, writing smart contracts, and integrating data verification functions. This process necessitates substantial investments in human, material, and financial resources while facing challenges related to immature technology and a lack of standardization. Additionally, blockchain solutions must integrate with the existing systems of brand owners, addressing discrepancies in system architecture, data standards, and data exchange protocols, which adds to the complexity and overall cost of R&D. In terms of hardware, blockchain solutions need to be deployed on servers, networks, and storage devices, requiring investments in equipment, physical space, and considerations of performance, security, and stability. Furthermore, the adoption of blockchain solutions necessitates the use of coding and reading devices for processing food information, which involves purchasing or leasing the necessary equipment while considering factors such as cost, efficiency, and compatibility. This paper explores the issues related to the supply chain operation management of food companies in the context of blockchain R&D and discusses the following questions:

(1) In the highly competitive food industry, what motivates members of the food supply chain to conduct blockchain R&D? Additionally, how does blockchain impact operational decisions within the food supply chain?
(2) When comparing blockchain independent R&D to collaborative R&D, which model is more advantageous for operational decisions? Furthermore, how do operational decisions in the food supply chain differ between two R&D scenarios?

(3) In a more complex food market environment (e.g., the asymmetric substitute foods), will the choice of blockchain R&D model change?

To the best of our knowledge, this paper is one of the pioneering analytical operations managements (OM) studies that explore operational decision-making strategies related to blockchain R&D, providing theoretical insights for food companies in selecting appropriate blockchain R&D models. Addressing the issue of poor information communication within food supply chains, we comprehensively consider the advantages of blockchain platforms in enhancing operational efficiency and reducing production costs. We conduct a comparative analysis of the operational decision-making differences between independent and collaborative R&D blockchain platform models and propose optimal choices for blockchain R&D models at different stages of development.

We organize the paper as follows. Section 2 provides a comprehensive review of related studies to identify gaps in the literature and appropriately position our contribution. In Section 3, we define the problem, outline its assumptions, and develop the model. Section 4 analyzes the value of blockchain R&D in the food supply chain. Section 5 further explores operational decisions for food supply chains under different Blockchain R&D models based on the context of asymmetric substitute foods and provide targeted management insights.

2. Literature review

2.1 Blockchain in supply chains

Numerous investigations have explored the potential advantages of distributed ledger technology for various sectors and enterprises within supply chain ecosystems [1, 2]. The implementation of this innovative technology in supply chain operations management is becoming increasingly prevalent. For instance, it has been applied to enhance supply chain transparency and verify product authenticity [3-6]. Furthermore, it has demonstrated its capacity to promote supply chain sustainability [7-9] and even revolutionize lean manufacturing practices within supply chains [10, 11]. Notably, the body of research examining the integration of distributed ledger technology with food supply chain management continues to expand rapidly.

The application of distributed ledger systems in supply chain management has seen a significant surge in interest. Numerous researchers have employed diverse methodologies to investigate the potential implementation of this technology in supply chain operations. For instance, Wu *et al.* took an analytical approach to examine strategies for adopting a distributed ledger technology system (DLTS) in fresh product supply chains (FPSC). They compared scenarios without this technology to three different situations where various FPSC members led the DLTS implementation. Their research yielded optimal conditions for DLTS deployment in FPSC and proposed a two-part tariff contract to coordinate DLTS construction [12]. Bumblauskas *et al.* utilized a case study approach to investigate the deployment of this technology in egg distribution. Their findings demonstrated how it can enhance accuracy and transparency in product movement throughout supply chains [13]. Kamble *et al.* adopted a hybrid methodology, combining Interpretive Structural Modeling and Decision-Making Trial and Evaluation Laboratory, to explore strategies for implementing distributed ledger systems in agricultural supply chains, with a focus on ensuring food safety and sustainability [14]. Wang *et al.* has examined blockchain's impact on pricing strategies in dual-channel supply chains, revealing how this technology incentivizes dynamic pricing adjustments across sales periods while accounting for strategic consumer behavior [15].

As the advantages of distributed ledger technology become increasingly apparent, researchers have begun to delve into specific operational challenges within this framework. For instance, Mangla *et al.* utilized system dynamics modeling to examine the implementation of distributed ledger technology in a sustainable dairy supply chain. This study showcased how technology could be leveraged to address sustainability concerns in a complex, perishable goods supply chain [16]. In the context of food supply networks, Rogerson and Parry employed case study methodologies to empirically demonstrate the capacity of distributed ledger systems to enhance supply chain visibility. Their research highlighted the transformative potential of this technology in improving transparency across the entire supply chain [17]. Taking a different approach, Behnke and Janssen conducted an empirical investigation into the requisite modifications of supply chain organizational structures and the importance of persuading supply chain participants to dismantle information silos [18]. Yang and Zhang proposed a blockchain-based production scheduling and control optimization (PSCO-PC) strategy for intelligent manufacturing. By introducing adaptive difficulty mechanisms and improving simulation model flexibility, the research addressed data throughput and consensus challenges. Experimental results validated the strategy's effectiveness in optimizing production resource control and enhancing matching rationality in manufacturing systems [19] .

Existing studies have explored blockchain technology adoption in various supply chains, primarily using methods like case analyses, empirical research, and interpretive structural modeling. However, game-theoretic analytical approaches remain scarce in this area. Our study differs by comparing operational decision-making between independent and collaborative blockchain R&D approaches. We also suggest optimal blockchain R&D strategies for different developmental stages. By comparing independent and collaborative blockchain R&D models, we provide theoretical insights that guide food companies in their strategic decision-making. Our analysis considers various factors, including food substitutability, quality credibility, and spillover effects, to determine optimal R&D levels and production strategies. This holistic approach allows us to propose tailored solutions for different supply chain scenarios, contributing to the overall effectiveness of food supply chains in an increasingly competitive landscape.

2.2 Competitive collaboration in supply chain

Research on competitive collaboration in supply chains has also gained significant attention. Various studies have explored different aspects of this phenomenon. For instance, an investigation into e-commerce channels examined optimal decisions and profits for online retailers and manufacturers across four service channel types, highlighting revenue sharing as a crucial factor [20]. In the fresh produce sector, a study on supplier competition revealed that freshness preservation efforts and retailer's information disclosure level significantly influence supply channel dynamics [21]. Another research focused on a manufacturer's service selection between competing module suppliers, suggesting that leveraging diverse pricing and service strategies from both suppliers could be more beneficial than relying on a single superior supplier [22]. The timing of pricing and marketing decisions in manufacturer-led supply chains has also been analyzed. Through a series of game-theoretic models, researchers identified optimal decision timing by comparing equilibrium outcomes across different supply chain configurations [23]. Furthermore, Deng employed machine learning-enhanced agent-based modeling to examine retailer price competition under consumer learning behavior and supplier competition. The study utilized fuzzy logic, genetic algorithms, reinforcement learning, and swarm intelligence to simulate market dynamics. Results showed that different consumer learning behaviors lead to varied retailer competition patterns, while supplier price competition affects the intensity of retailer price competition. The study provides a simulated market model for future research on price competition among supply chain actors [24].

In the context of blockchain application, Liu *et al*. investigated blockchain service provision in supply chains with downstream competition. Using game theory, they analyzed the optimal strategies for a manufacturer and two competing retailers. Their work explored how blockchain impacts market dynamics, economic outcomes, and service performance in competitive supply chain settings. These studies underscore the potential of blockchain to reshape competitive collaboration in modern supply chains [25]. Similarly, Song *et al*. examined how blockchain affects information sharing decisions among rival e-commerce sellers. Their research on a two-competitor market revealed that blockchain adoption becomes universal when consumer trust in information is low or implementation costs are minimal. This highlights how blockchain can foster collaboration even in competitive environments [26]. Yan *et al*. compared blockchain-based and traditional approaches to supply chain information coordination. They developed a three-level supply chain model incorporating retailer information sensitivity. Their study revealed that blockchain technology can effectively reduce operating costs. Interestingly, they found that moderate levels of information-sensitive retailers optimize blockchain value, as extreme levels may increase privacy concerns among supply chain companies [27].

Our research builds upon previous studies by offering a comprehensive analysis of blockchain's role in food supply chains, focusing on operational efficiency and production costs. Unlike earlier work that primarily examined competitive dynamics and information disclosure, we investigate the critical choice between independent and collaborative blockchain R&D approaches. This study addresses the unique challenges in food supply chains, such as consumer trust and collaborative production costs.

3. The model

This paper considers the operational decision-making issues of two food companies in the context of a duopoly market. Food companies i $(i = 1, 2)$ each sell similar types of food with a certain degree of substitutability. As blockchain technology begins to be applied in various food segments, both companies are aware that blockchain R&D can enhance market demand potential, increase consumer trust in food quality, and reduce food production costs. The food production and sales price of company are denoted as q_i and p_i, respectively, while the blockchain R&D effort is denoted as s_i, the higher intensity of blockchain R&D, the more production costs the company can save, and consumer trust in the food will also increase. To improve the readability, Table 1 summarizes all abbreviations and definitions of important variables involved.

Table 1 Notation used in this paper

Notation	Definition
q_i	Production of food company i
s_i	The blockchain R&D effort level(the level of blockchain services)
p_i	The sales price of food company i
π_i	The profit of food company i
Q	Total production of the food supply chain
θ	Market demand potential under blockchain adoption
α	Market demand potential without blockchain adoption
β	Food substitutability coefficient
c	Unit production cost of food companies without blockchain adoption
C_i	Total production cost of food companies with blockchain adoption
γ	The coefficient of blockchain R&D spillover effect
x	Food quality
t	The credibility of food's quality

3.1 Without considering blockchain R&D

As the benchmark model, we first consider the scenario where blockchain technology is conducted in the food supply chain. Without considering the blockchain R&D, the two food companies do not engage in collaborative R&D. Followed the previous studies [28-31], the respective inverse demand functions of can be expressed as follows:

$$p_i^{NN} = \alpha - q_i^{NN} - \beta q_{3-i}^{NN} + xt \tag{1}$$

where α represents the market demand potential without blockchain technology. The market demand potential is normalized to 1 $(\alpha = 1)$. This assumption is consistent with the research hypothesis of Niu et al. (2021) and does not affect the main results. β represents the food substitutability coefficient $(0 < \beta < 1)$; x denotes the food quality. Premium food typically refers to products of superior quality, high nutritional value, and fine processing, such as organic or imported foods. Ordinary food refers to products of average quality and lower processing levels, like everyday groceries; t represents consumer trust in food quality $(0 \leq t < 1)$. When $t = 0$, it indicates that consumers have no trust in the information described on food packaging and completely distrust the stated food quality. In this case, only consumers who are entirely insensitive to food quality will choose to purchase, deciding solely based on price. Clearly, as consumer trust in food quality increases, they will be willing to pay a higher price for the food. In the blockchain context, t also reflects consumer adoption of verification tools - higher t implies not only greater trust but also more consumers actively using blockchain verification, which directly enhances the market value of blockchain investment.

The profit functions for the two food companies can be expressed as follows:

$$\pi_i^{NN} = p_i^{NN} q_i - c q_i^{NN} = (\alpha - q_i^{NN} - \beta q_{3-i}^{NN} + xt - c) q_i^{NN} \tag{2}$$

where c represents the unit production cost of food ($0 < c < \alpha + xt$). Under equilibrium conditions, the optimal quantity for each food company and the total production of the food supply chain can be expressed as:

$$q_i^{NN*} = \frac{1 + tx - c}{2 + \beta} \tag{3}$$

$$Q^{NN*} = \frac{2(1 + tx - c)}{2 + \beta} \tag{4}$$

The optimal profit for each food company can be expressed as:

$$\pi_i^{NN*} = \frac{(1 - c + tx)^2}{(2 + \beta)^2} \tag{5}$$

3.2 Independent blockchain R&D

When both food companies decide to develop blockchain services, firstly, the addition of blockchain technology to food products will attract more consumers, thereby increasing the potential demand for such products in the market. Secondly, because information about the entire process from production to distribution will be recorded and verified in their respective blockchain services, consumer trust in the quality and safety of the food will be strengthened. Additionally, the reduction in food production waste and improvement in production efficiency will lower the production costs for the companies. Therefore, when food companies choose to conduct blockchain R&D independently, the following applies:

$$p_i^{NB} = \theta - q_i^{NB} - \beta q_{3-i}^{NB} + x(t + s_i^{NB}) \tag{6}$$

$$C^{NB} = \left(c - s_i^{NB} - \gamma s_{3-i}^{NB}\right) q_i^{NB} \tag{7}$$

$$\pi_i^{NB} = p_i^{NB} q_i^{NB} - C^{NB} - \frac{{s_i^{NB}}^2}{2} \tag{8}$$

In the above formula, γ represents the spillover effect of blockchain R&D ($0 < \gamma < 1$). As an innovation activity that enhances production efficiency, the blockchain R&D may have positive spillover effects on other companies in the market that undertake similar R&D actions. For the production process, blockchain technology greatly ensures the credibility and security of food production and distribution data, and the adoption of blockchain can reduce the data verification and audit costs for both parties involved in the transaction. For the transaction process, if both food companies use smart contract features of blockchain technology to simplify and automate market transactions, the risk of transaction defaults and time costs will decrease.

From an industry development perspective, blockchain R&D by any food company will drive the digital transformation of the entire industry supply chain, promoting improvements in production efficiency across the industry. Therefore, it can be inferred that the external effects of a food company's R&D activities will lower the unit production costs of its competitors.

It is also worth noting that, in practice, regulatory and data-sharing constraints significantly influence the magnitude of spillover effects. In regions with relaxed regulatory environments, inter-firm technology exchange and data sharing are more convenient, resulting in relatively higher γ values. However, in jurisdictions with strict data protection regulations, information sharing between companies faces more legal restrictions, leading to lower γ values. Thus, γ reflects not only technological knowledge diffusion but also the impact of institutional environments on collaborative R&D feasibility.

Additionally, the relationship between R&D expenditure and returns is modeled using a quadratic function to reflect the diminishing returns of R&D investment. Thus, given s_i^{NB}, the equilibrium production quantities for the two food companies can be expressed as:

$$\begin{cases} q_1^{NB}(s_1^{NB},s_2^{NB})^* = \frac{\theta+tx-c}{2+\beta} + \frac{s_1^{NB}(2+2x-\beta\gamma)}{4-\beta^2} + \frac{s_2^{NB}(2\gamma-\beta-x\beta)}{4-\beta^2} \\ q_2^{NB}(s_1^{NB},s_2^{NB})^* = \frac{\theta+tx-c}{2+\beta} + \frac{s_1^{NB}(2\gamma-\beta-x\beta)}{4-\beta^2} + \frac{s_2^{NB}(2+2x-\beta\gamma)}{4-\beta^2} \end{cases} \tag{9}$$

Based on this, the total market demand for food in the production phase can be expressed as:

$$\begin{aligned} Q^{NB}(s_1^{NB},s_2^{NB})^* &= \frac{2\theta+tx-c}{2+\beta} + \frac{s_1^{NB}(2-\beta)(1+x+\gamma)}{4-\beta^2} + \frac{s_2^{NB}(2-\beta)(1+x+\gamma)}{4-\beta^2} \\ &= \frac{2R}{2+\beta} + \frac{Y}{2+\beta} \cdot 2s_i^{NB} \end{aligned} \tag{10}$$

For simplicity in the formula, let $Y = 1 + x + \gamma$, which can be understood as the positive impact of the R&D spillover effect on the level of R&D; let $R = \theta + tx - c$, which can be understood as the marginal revenue per unit of food. In the expression for total food market demand in the production phase, the first term $2R/(2+\beta)$ represents the sum of the equilibrium quantities of the two food companies without R&D; the second term $Y/(2+\beta)$ represents the positive impact factor of R&D level on total demand. The numerator reflects the impact of the R&D spillover effect, with a larger γ indicating a stronger spillover effect and thereby increasing total demand. The denominator β reflects the sensitivity of market demand to food substitutability, with a larger β indicating more intense market homogenization and thus reducing total demand.

By substituting $q_1^{NB}(s_1^{NB},s_2^{NB})^*$and $q_2^{NB}(s_1^{NB},s_2^{NB})^*$ into π_1^{NB} and π_2^{NB}, and calculating the second derivative of the profit expressions for the two food companies, it can be determined that maximum profit occurs when $(2+2x-\beta\gamma)/(4-\beta^2) < \sqrt{2}/2$. Thus, the optimal R&D level for food companies under independent blockchain R&D is:

$$s_i^{NB^*} = \frac{2NR}{BB' - 2NY} \tag{11}$$

where $N = 2 + 2x - \beta\gamma$, which can be understood as the negative impact of the R&D spillover effect on the level of R&D; NY can represent the complementarity between the R&D levels of the two companies. When the R&D spillover effect γ is large, this value is smaller, indicating that the R&D levels of the two companies mutually enhance each other, thus increasing total demand. When γ is small, this value is larger, indicating that the R&D levels of the two companies counteract each other, thus reducing total demand; $B = 4 - \beta^2$, $B' = 2 + \beta$, and $B \cdot B'$ can be understood as the negative impact of food substitutability on the level of R&D.

Under independent blockchain R&D model, the optimal food production quantities for each food company and the total food production can be expressed as:

$$q_i^{NB^*} = \frac{BR}{BB' - 2NY} \tag{12}$$

$$Q^{NB^*} = \frac{2BR}{BB' - 2NY} \tag{13}$$

The optimal profit for each food company can be expressed as:

$$\pi_i^{NB} = \frac{R^2(B^2 - 2N^2)}{(BB' - 2NY)^2} \tag{14}$$

3.3 Collaborative blockchain R&D

This model examines the scenario in which two food companies engage in collaborative efforts during the blockchain R&D phase. When these companies opt for cooperative R&D initiatives, the extent and sophistication of blockchain R&D are collectively determined through the combined efforts and resources of both companies, so $s^{CRB} = s_1^{CRB} = s_2^{CRB}$.

First, using the same calculation methods as in the previous section, we calculate the optimal profit for each food company. Accordingly, under blockchain R&D collaboration, the joint profit function for the two food companies can be expressed as:

$$\pi^{CRB} = \pi_1^{CRB} + \pi_2^{CRB} \tag{15}$$

Accordingly, in the production phase, given s^{CRB}, the equilibrium production quantities for the two food companies can be expressed as:

$$q_i^{CRB}(s^{CRB})^* = \frac{R + Ys^{CRB}}{2+\beta} = \frac{2R}{2+\beta} + \frac{Y}{2+\beta} \cdot 2s_i^{NB} \tag{16}$$

When $Y/(2+\beta) < 1/\sqrt{2}$, the food companies achieve maximum profit. Thus, the optimal R&D level for the two food companies under independent blockchain R&D is:

$$s^{CRB^*} = \frac{2RY}{B'^2 - 2Y^2} \tag{17}$$

$$q_i^{CRB^*} = \frac{RB'}{B'^2 - 2Y^2} \tag{18}$$

$$\pi_i^{CRB^*} = \frac{R^2}{B'^2 - 2Y^2} \tag{19}$$

4. The value of blockchain R&D in food supply chain

By comparing the equilibrium quantities under three scenarios—no blockchain adoption, independent blockchain R&D, and collaborative blockchain R&D, the following conclusions can be drawn:

Proposition 1: (1) $q_i^{NN^*} < q_i^{NB^*}$, $q_i^{NN^*} < q_i^{CRB^*}$; (2) when $2\gamma - (1+x)\beta > 0$, then $q_i^{NB^*} < q_i^{CRB^*}$; when $2\gamma - (1+x)\beta < 0$, then $q_i^{NB^*} > q_i^{CRB^*}$.

Proposition 1 indicates that (1) The food supply chain can improve its optimal production quantity by conducting blockchain R&D. (2) The impact of the blockchain R&D model on the equilibrium production quantity of the food supply chain is influenced jointly by the blockchain R&D spillover effect (γ), food quality (x), and food substitutability (β).

To illustrate the findings of Proposition 1 more clearly, we perform calculations with assigned values for the relevant variables. Assume $\theta = 1.2$, $c = 0.8$, $t = 0.9$. The effect of variations in γ, x, and β within their respective ranges on the difference in food supply chain production quantities is shown in Fig. 1.

Firstly, by comparing the feasible region where $q_i^{CRB^*} > q_i^{NB^*}$, it is observed that the independent R&D model is more likely to achieve higher equilibrium food production quantities compared to the collaborative R&D model.

Secondly, as food quality increases, the region where collaborative R&D blockchain can achieve higher food production quantities becomes increasingly constrained. This suggests that high-quality food is typically a key product for food companies with high commercial value. If the two food companies collaborate on blockchain R&D, the sales data and flow information of high-quality food are likely to be acquired by competitors, leading to a gradual reduction in their optimal production quantity. However, if the two food companies develop blockchain R&D independently, each can maintain its own data and information, better protect its commercial secrets and competitive advantage, and adjust its production and sales strategies more flexibly to meet consumer demands and preferences, thereby increasing its market share.

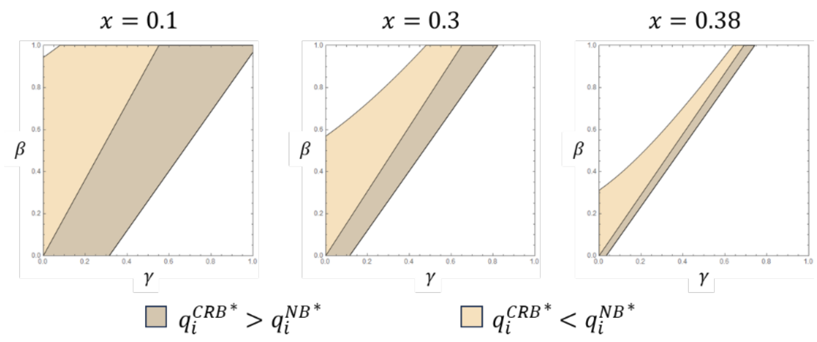


Fig. 1 Evolution of the dominant region between $q_i^{NB^*}$ and $q_i^{CRB^*}$

Proposition 2: (1) when $\theta > \theta'$, $\pi_i^{NN^*} < \pi_i^{NB^*}$; (2) when $\theta < \theta'$, $\pi_i^{NN^*} > \pi_i^{NB^*}$; (2) $\pi_i^{NN^*} < \pi_i^{CRB^*}$, $\pi_i^{NB^*} < \pi_i^{CRB^*}$.

Proposition 2 indicates that the adoption of blockchain does not necessarily lead to profit increases for members of the food supply chain. When food companies choose to independently conduct blockchain R&D, it is only more profitable for food companies to opt for independent blockchain R&D if the blockchain significantly enhances the potential demand for the food supply chain. Assuming $x = 0.2$ and keeping the other parameters constant, Fig. 2 illustrates the range of profit differences between choosing to develop and not to conduct blockchain R&D in the food supply chain.

It can be observed that, firstly, under different settings of potential market demand for food, the area where $\pi_i^{NN^*} < \pi_i^{NB^*}$ is always larger than the area where $\pi_i^{NN^*} > \pi_i^{NB^*}$. This suggests that, although there are instances where not conducting blockchain R&D yields higher profits, choosing independent R&D is more likely to result in greater economic benefits. Secondly, as the blockchain increases the potential market demand for food, represented by θ, the feasible area for $\pi_i^{NN^*} < \pi_i^{NB^*}$ expands accordingly, indicating that the economic incentive for food companies to conduct blockchain R&D also strengthens. While our model treats consumer trust t as a constant parameter, in reality, trust in blockchain systems can be undermined by misinformation or security breaches. Such trust erosion would effectively reduce t, thereby raising the profitability threshold θ in Proposition 2. This means even greater market demand enhancement would be required to justify blockchain investment. For instance, a security breach that reduces t could shift a company from the profitable region (where $\pi_i^{NN^*} < \pi_i^{NB^*}$) to the unprofitable region, highlighting why maintaining blockchain integrity is crucial for sustained profitability.

However, if food companies choose to collaborate on blockchain R&D, the profit level of the food supply chain can be significantly enhanced compared to the other two blockchain development models mentioned above. Thus, choosing collaborative R&D is a more economically viable operational strategy.

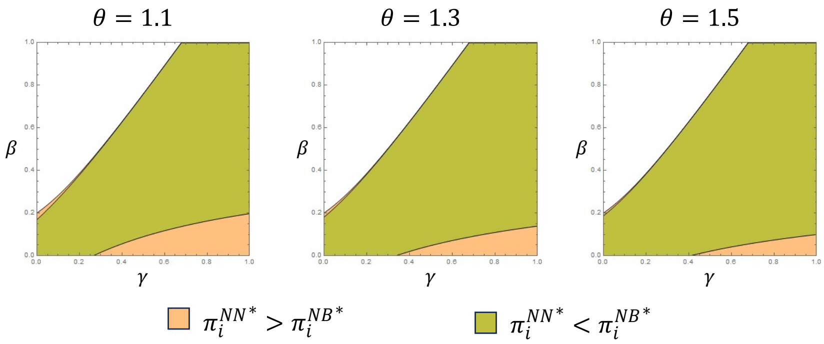


Fig. 2 Evolution of the dominant region between $\pi_i^{NN^*}$ and $\pi_i^{NB^*}$

Proposition 3: (1) when $\gamma < \frac{(1+x)\beta}{2}$, $s_i^{CRB^*} > s_i^{NB^*}$; (2) when $\gamma > \frac{(1+x)\beta}{2}$, $s_i^{CRB^*} < s_i^{NB^*}$.

Proposition 3 indicates that when the spillover effect level of blockchain R&D is below a certain threshold, a collaborative R&D model enables the food supply chain to achieve a higher level of R&D effort. Conversely, when the spillover effect level is higher, an independent R&D model can achieve a higher level of R&D effort. Specifically, when the blockchain R&D spillover effect is relatively low, it implies that the application of blockchain technology is not yet widespread and deep enough. In such cases, deeper integration with other technologies like the IoT, big data, and AI is required, which increases the complexity and cost of technology development and maintenance. If a company opts for a collaborative R&D model, it can coordinate and integrate the technical resources and advantages of other companies within the supply chain, thus reducing the difficulty and risk of technology development and maintenance. This approach can help achieve a higher level of blockchain R&D in the food supply chain. On the other hand, when the blockchain R&D spillover effect level is high, it indicates that blockchain technology is already quite mature and stable. In such a scenario, food companies can customize the blockchain services according to the specific characteristics of their supply chain, achieving a higher level of blockchain service application. This not only helps to protect their trade secrets and competitive advantages but also enhances their market competitiveness.

Fig. 3 illustrates the variation in the difference between the level of blockchain collaborative R&D and independent R&D under different levels of R&D spillover effects. As γ increases, the feasible region where the collaborative R&D model achieves a higher R&D level gradually expands and moves towards the upper left region of the graph. This suggests that when the spillover effect of R&D is low, food companies lack sufficient motivation and resources to engage in R&D collaboration. However, as the spillover effect of R&D increases, the production efficiency of food companies is significantly enhanced with the help of blockchain services, prompting them to have a stronger incentive to engage in collaborative R&D and improve the level of blockchain services. On the other hand, this move may also intensify competition in the food market (as shown in the graph by the increase in the food substitutability rate, β). Furthermore, implementation time lags in blockchain deployment create additional strategic considerations, as early collaborative investors must balance the immediate costs against delayed spillover benefits, potentially moderating the attractiveness of collaboration even in high-spillover environments.

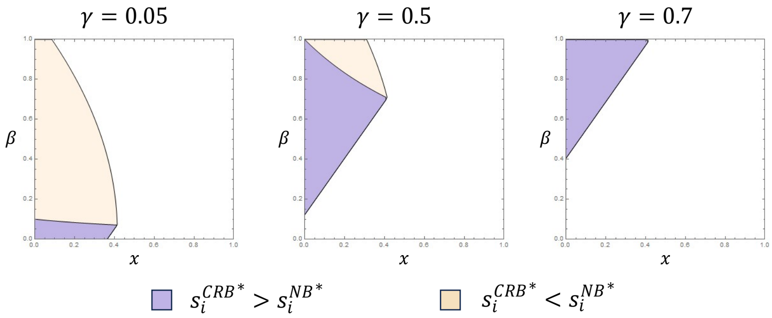


Fig. 3 Evolution of the dominant region between $s_i^{CRB^*}$ and $s_i^{NB^*}$

5. Extended analysis

5.1 Asymmetric substitute foods

This section expands the research background to include asymmetric substitute food supply chains, considering supply chain operational decisions within the more complex context of blockchain R&D collaboration. Asymmetric substitute foods refer to cases in the food market where two or more types of food exist, and their substitutability is not mutual but rather one-directional. For example, soy milk can serve as a substitute for cow's milk, catering to the needs of individuals who are lactose intolerant or follow a vegetarian diet. However, cow's milk cannot completely replace the nutritional value and health benefits of soy milk. Based on these industry observations, this subsection assumes that the food product of Company 1 is superior to that of Company 2. Thus, the former can be considered a perfect substitute for the latter, but not vice versa. In the remaining content of this subsection, the product of Company 1 will be referred to as the "superior food", while the product of Company 2 will be referred to as the "inferior food". To simplify the modeling process, this subsection assumes that β_2=1 and β_1=β<1. Therefore, under independent R&D, the inverse demand function for the food supply chain can be rewritten as follows:

$$p_1^{ANB} = \theta - q_1^{ANB} - \beta q_2^{ANB} + x(t + s_1^{ANB}) \tag{20}$$

$$p_2^{ANB} = \theta - q_2^{ANB} - q_1^{ANB} + x(t + s_2^{ANB}) \tag{21}$$

Furthermore, under collaborative R&D, the inverse demand function for the food supply chain can be rewritten as follows:

$$p_1^{ACB} = \theta - q_1^{ACB} - \beta q_2^{ACB} + x(t + s_1^{ACB}) \tag{22}$$

$$p_2^{ACB} = \theta - q_2^{ACB} - q_1^{ACB} + x(t + s_2^{ACB}) \tag{23}$$

Referring to similar calculation steps from the previous text, the equilibrium decision results for the two food companies in the asymmetric substitute food supply chain, under both independent and collaborative blockchain R&D scenarios, can be easily derived.

5.2 Comparison of R&D levels in asymmetric substitute food companies

By comparing the equilibrium R&D levels of asymmetric substitute food companies under independent blockchain R&D model, the following conclusions can be drawn:

Proposition 4: when $\beta > 4(\sqrt{2} - 1 - x)/(\sqrt{2} - 2\gamma)$ or $\beta < 4(\sqrt{2} - 1 - x)/(\sqrt{2} - 2\gamma)$ and $3\sqrt{2} + 2\gamma \geq 4 + 4x$, $s_1^{ANB^*} < s_2^{ANB^*}$; Otherwise $s_1^{ANB^*} > s_2^{ANB^*}$.

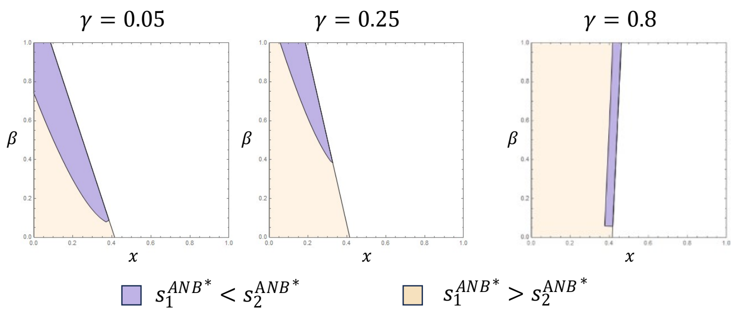


Fig. 4 Evolution of the dominant region between $s_1^{ANB^*}$ and $s_2^{ANB^*}$

Proposition 4 indicates that the relative sizes of food quality levels, food substitutability rates, and blockchain R&D spillover effects collectively determine the R&D levels of the two food companies. When R&D spillover effects are low, the inferior food company may have a higher R&D level than the superior food company under lower market competition or higher food quality levels. This suggests that, under lower market competition, superior food companies may lack the motivation to invest in blockchain R&D, as they already hold a significant market share. In contrast, inferior food companies need to invest in blockchain R&D to enhance their market position and increase their revenue and profit. At higher food quality levels, superior food companies already possess high quality and reputation, so there is less need to invest in blockchain services to further enhance their brand image. Meanwhile, inferior food companies need to use blockchain services to demonstrate that their products are also of high quality and reliability, thereby attracting more high-end consumers.

Using the same parameter settings as in the previous subsection, and setting γ to $\gamma = 0.05, 0.25, 0.8$ respectively, Fig. 4 illustrates the comparative results of R&D levels for asymmetric substitute food companies under independent blockchain R&D. The figure shows that, under various R&D spillover parameter settings, the feasible domain where $s_1^{ANB^*} > s_2^{ANB^*}$ is significantly larger than the domain where $s_1^{ANB^*} < s_2^{ANB^*}$. This indicates that superior food companies are more likely to set higher blockchain R&D levels to maintain their dominant position in the market.

5.3 Value of R&D collaboration for asymmetric substitute food supply chains

By comparing the optimal production quantities and optimal profits of the two food companies in the context of independent blockchain R&D for asymmetric substitute foods, the following conclusions can be drawn:

Proposition 5: (1) $Q^{ANB^*} < Q^{ACB^*}$; (2) $\pi^{ANB^*} < \pi^{ACB^*}$.
Proposition 5 indicates that in an asymmetric substitute food supply chain, collaborative blockchain R&D can lead to higher overall consumer demand and total profit in the food market. Comparing this with the conclusions from Proposition 1, it is evident that, when food substitutability is asymmetric, the food production under collaborative R&D will be strictly higher than under independent R&D.

For members of the food supply chain, this asymmetric substitutability means that the impact of collaborative R&D on blockchain services differs between companies. For Company 1, collaborative R&D can enhance the advantages of its food products, thereby consolidating its market leadership, capturing a larger market share, and increasing its profit levels. For Company 2, collaborative R&D can help compensate for the disadvantages of its food products, thereby improving its market competitiveness, attracting more consumers, and boosting its profit levels.

6. Conclusion

This paper explores the blockchain R&D decision-making issues faced by food companies in the context of the increasing application of blockchain technology in food supply chain management. It focuses on two competing food companies simultaneously developing blockchain services on their self-built blockchain platforms. The paper examines the motivations for developing blockchain services, the impact of blockchain services on supply chain operational decisions, differences in supply chain operational decisions under independent and collaborative R&D models, and changes in R&D model choices in more complex food market environments (such as asymmetric substitute foods). To address these issues, the paper establishes several models: Supply chain operational decisions without blockchain R&D collaboration (Benchmark Model), supply chain operational decisions with collaborative blockchain R&D (Collaborative R&D Model), and supply chain operational decisions under the context of asymmetric substitute foods (Asymmetric Substitute Foods Model).

By comparing optimal operational decisions for food companies under three models—no blockchain R&D, independent R&D, and collaborative R&D—the paper derives the optimal model and operational strategy for blockchain R&D collaboration in food supply chains. Additionally, the paper revisits operational decision schemes under different R&D models in the context of asymmetric substitute foods to test the robustness of the benchmark model results and provide more targeted management insights:

From the perspective of food supply chain production, (1) Independent blockchain R&D tends to achieve higher market equilibrium production compared to collaborative R&D. This is because independent R&D better protects food companies' trade secrets and competitive advantages, whereas collaborative R&D may lead to information leakage and imitation by competitors. (2) High-value premium food supply chains are more suited for independent blockchain R&D, while ordinary food supply chains are better suited for collaborative R&D. Premium foods have higher commercial value and competitive advantages, giving supply chain members a stronger incentive and ability to protect their trade secrets and intellectual property, thus reducing the risk of information leakage and imitation. Ordinary foods, being more homogeneous, benefit more from collaboration to reduce R&D costs and risks.

From the perspective of food supply chain profit: The application of blockchain does not necessarily increase profits for food supply chain members. For independent R&D, it is only profitable when the blockchain's potential demand enhancement effect reaches a certain level. For collaborative R&D, the profit level of the food supply chain can be significantly higher compared to the other two blockchain R&D models. Therefore, choosing collaborative R&D is a more economical operational strategy.

From the perspective of blockchain R&D, both independent and collaborative model of blockchain R&D can achieve relatively higher equilibrium R&D levels. The level of blockchain R&D spillover effects reflects the stage of blockchain service application in the food market. When blockchain technology is not yet widely and deeply applied, the R&D spillover effect level is lower, making collaborative R&D a better choice. As the adoption of blockchain and R&D progress, the spillover effect level increases, and food companies will gradually shift towards independent R&D model to achieve higher levels of blockchain service and stronger market competitiveness. Additionally, when considering further asymmetric food R&D contexts, it is found that the R&D levels of blockchain under independent and collaborative models are also influenced by the level of food differentiation and food quality. Inferior goods may set higher R&D levels in independent R&D to prove their food quality and attributes to consumers.

We propose several avenues for future research. First, exploring how consumer trust in blockchain affects the adoption of independent versus collaborative R&D models could provide valuable insights into strategic decision-making in food supply chains. Second, future studies should investigate the role of blockchain R&D in enhancing sustainability practices within food supply chains, particularly regarding transparency and traceability. Lastly, examining the long-term impact of blockchain R&D on competitive advantage will be crucial as companies transition from collaborative to independent models in response to market dynamics.

Acknowledgement

This work was supported by the [National Natural Science Foundation of China] under Grant [72401283].

References

[1] Gurtu, A., Johny, J. (2019). Potential of blockchain technology in supply chain management: A literature review, *International Journal of Physical Distribution & Logistics Management*, Vol. 49, No. 9, 881-900, doi: 10.1108/IJPDLM-11-2018-0371.

[2] Wang, Y., Singgih, M., Wang, J., Rit, M. (2019). Making sense of blockchain technology: How will it transform supply chains?, *International Journal of Production Economics*, Vol. 211, 221-236, doi: 10.1016/j.ijpe.2019.02.002.

[3] Choi, T.-M. (2019). Blockchain-technology-supported platforms for diamond authentication and certification in luxury supply chains, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 128, 17-29, doi: 10.1016/j.tre.2019.05.011.

[4] Čuš-Babič, N., Rebolj, D., Nekrep-Perc, M., Podbreznik, P. (2014). Supply-chain transparency within industrialized construction projects, *Computers in Industry*, Vol. 65, No. 2, 345-353, doi: 10.1016/j.compind.2013.12.003.

[5] Pun, H., Swaminathan, J.M., Hou, P. (2021). Blockchain adoption for combating deceptive counterfeits, *Production and Operations Management*, Vol. 30, No. 4, 864-882, doi: 10.1111/poms.13348.

[6] Sunny, J., Undralla, N., Pillai, V.M. (2020). Supply chain transparency through blockchain-based traceability: An overview with demonstration, *Computers & Industrial Engineering*, Vol. 150, Article No. 106895, doi: 10.1016/j.cie.2020.106895.

[7] Bai, C., Sarkis, J. (2020). A supply chain transparency and sustainability technology appraisal model for blockchain technology, *International Journal of Production Research*, Vol. 58, No. 7, 2142-2162, doi: 10.1080/00207543.2019.1708989.

[8] Saberi, S., Kouhizadeh, M., Sarkis, J., Shen, L. (2019). Blockchain technology and its relationships to sustainable supply chain management, *International Journal of Production Research*, Vol. 57, No. 7, 2117-2135, doi: 10.1080/00207543.2018.1533261.

[9] Dong, Y., Song, L., Fu, L., Zhang, G. (2024). Humanitarian supply chain agility: A blockchain integration perspective, *Tehnički Vjesnik – Technical Gazette*, Vol. 31, No. 5, 1660-1667, doi: 10.17559/TV-20230425000568.

[10] Chang, J.A., Katehakis, M.N., Shi, J.J., Yan, Z. (2021). Blockchain-empowered newsvendor optimization, *International Journal of Production Economics*, Vol. 238, Article No. 108144, doi: 10.1016/j.ijpe.2021.108144.

[11] Choi, T.-M., Luo, S. (2019). Data quality challenges for sustainable fashion supply chain operations in emerging markets: Roles of blockchain, government sponsors and environment taxes, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 131, 139-152, doi: 10.1016/j.tre.2019.09.019.

[12] Wu, X., Fan, Z., Cao, B. (2021). An analysis of strategies for adopting blockchain technology in the fresh product supply chain, *International Journal of Production Research*, Vol. 61, No. 11, 3717-3734, doi: 10.1080/00207543.2021.1894497.

[13] Bumblauskas, D., Mann, A., Dugan, B., Rittmer, J. (2020). A blockchain use case in food distribution: Do you know where your food has been?, *International Journal of Information Management*, Vol. 52, Article No. 102008, doi: 10.1016/j.ijinfomgt.2019.09.004.

[14] Kamble, S.S., Gunasekaran, A., Sharma, R. (2020). Modeling the blockchain enabled traceability in agriculture supply chain, *International Journal of Information Management*, Vol. 52, Article No. 101967, doi: 10.1016/j.ijinfomgt.2019.05.023.

[15] Wang, G., Cao, E., Weng, J., Xie, S. (2024). Dynamic pricing in a blockchain-enabled dual-channel supply chain, *Tehnički Vjesnik – Technical Gazette*, Vol. 31, No. 3, 753-773, doi: 10.17559/TV-20230726000830.

[16] Mangla, S.K., Kazancoglu, Y., Ekinci, E., Liu, M., Özbiltekin, M., Sezer, M.D. (2021. Using system dynamics to analyze the societal impacts of blockchain technology in milk supply chainsrefer, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 149, Article No. 102289, doi: 10.1016/j.tre.2021.102289.

[17] Rogerson, M., Parry, G.C. (2020). Blockchain: Case studies in food supply chain visibility, *Supply Chain Management: An International Journal*, Vol. 25, No. 5, 601-614, doi: 10.1108/SCM-08-2019-0300.

[18] Behnke, K., Janssen, M.F.W.H.A. (2020). Boundary conditions for traceability in food supply chains using blockchain technology, *International Journal of Information Management*, Vol. 52, Article No. 101969, doi: 10.1016/j.ijinfomgt.2019.05.025.

[19] Yang, S.Y., Zhang, M.F. (2023). Blockchain-driven optimization in intelligent manufacturing, *International Journal of Simulation Modelling*, Vol. 22, No. 2, 350-361, doi: 10.2507/IJSIMM22-2-CO10.

[20] Shen, J., Qian, J., Shao, X. (2022). Service channel choice for supply chain under online marketplace, *International Transactions in Operational Research*, Vol. 29, No. 4, 2401-2422, doi: 10.1111/itor.13095.

[21] Chen, Z., Dan, B., Ma, S., Tian, Y. (2022). Demand information sharing of fresh produce supply chain considering competing suppliers' freshness-keeping effort, *International Transactions in Operational Research*, Vol. 31, No. 2, 1206-1231, doi: 10.1111/itor.13177.

[22] Liu, L., Zhao, L., Lee, C. (2021). Competing product service in an OEM supply chain with upstream module suppliers as service providers, *International Transactions in Operational Research*, Vol. 28, No. 6, 3364-3393, doi: 10.1111/itor.12980.

[23] Karray, S., Martín-Herrán, G. (2021). The effects of decision timing for pricing and marketing efforts in a supply chain with competing manufacturers, *International Transactions in Operational Research*, Vol. 31, No. 1, 568-615, doi: 10.1111/itor.13094.

[24] Deng, G.F. (2023). Dynamic price competition market for retailers in the context of consumer learning behavior and supplier competition: Machine learning-enhanced agent-based modeling and simulation, *Advances in Production Engineering & Management*, Vol. 18, No. 4, 434-446, doi: 10.14743/apem2023.4.483.

[25] Liu, J., Zhao, H., Lyu, Y., Yue, X. (2022). The provision strategy of blockchain service under the supply chain with downstream competition, *Annals of Operations Research*, Vol. 327, 375-400, doi: 10.1007/s10479-022-05034-2.

[26] Song, Y., Liu, J., Zhang, W., Li, J. (2022). Blockchain's role in e-commerce sellers' decision-making on information disclosure under competition, *Annals of Operations Research*, Vol. 329, 1009-1048, doi: 10.1007/s10479-021-04276-w.

[27] Yan, K., Cui, L., Zhang, H., Liu, S., Zuo, M. (2022). Supply chain information coordination based on blockchain technology: A comparative study with the traditional approach, *Advances in Production Engineering & Management*, Vol. 17, No. 1, 5-15, doi: 10.14743/apem2022.1.417.

[28] Yang, S.Y., Tan, C. (2022). Blockchain-based collaborative management of job shop supply chain, *International Journal of Simulation Modelling*, Vol. 21, No. 2, 364-374, doi: 10.2507/IJSIMM21-2-CO10.

[29] Hu, H., Li, Y., Li, M. (2024). Dual-channel supply chain coordination considering the consumer's perception of quality, *Economic Computation and Economic Cybernetics Studies and Research*, Vol. 58, No. 2, 198-216, doi: 10.24818/18423264/58.2.24.12.

[30] Wang, Y., Yang, L., Chen, J.H., Li, P. (2023). Supply chain game analysis based on mean-variance and price risk aversion under different power structures, *Advances in Production Engineering & Management*, Vol. 18, No. 1, 104-115, doi: 10.14743/apem2023.1.460.

[31] Wang, S.L., Zhang, Y., Sheng, X., Luo, X.Y. (2023). Blockchain in supply chain collaboration: A quantitative study, *International Journal of Simulation Modelling*, Vol. 22, No. 3, 532-543, doi: 10.2507/IJSIMM22-3-CO15.

APEM journal
Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 75–86
https://doi.org/10.14743/apem2025.1.528
ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Analysis and optimization of micro-milling parameters for improving part quality in ultrafine graphite with varying workpiece inclination angles

Kramar, D.[a,*], **Miljuskovic, G.**[a], **Cica, Dj.**[b]

[a]University of Ljubljana, Faculty of Mechanical Engineering, Slovenia
[b]University of Banja Luka, Faculty of Mechanical Engineering, Bosnia and Herzegovina

A B S T R A C T

Micro-milling is recognized as one of the most important manufacturing technologies for producing micro-components/products. Amongst various materials, graphite has an important role in conventional micro-electrical discharge machining electrodes. This paper is focused on the investigation of the effect of micro-milling process parameters on the dimensional accuracy and surface quality of ultrafine grain graphite TTK-4. Depth of cut, spindle speed, stepover distance and feed rate have been considered as process variables of micro ball-end milling in experimental design. Moreover, the influence of the workpiece's inclination angle was also investigated. Taguchi's L_9 (3^4) orthogonal array was chosen to design the experiments, whereas grey relational analysis (GRA) was utilized for the multi-objective optimization of the micro ball end milling process with minimum dimensional deviation and minimum arithmetic mean roughness as objective functions. Furthermore, principal component analysis (PCA) was used to extract principal components and identify the corresponding weights for performance characteristics. In order to determine the significance of micro-milling parameters on overall machining performance, analysis of variance (ANOVA) was performed. The result of the study revealed that the proposed approach is adequate to address the multi-objective optimization of micro-milling parameters.

A R T I C L E I N F O

Keywords:
Micro-milling;
Graphite;
Workpiece inclination angle;
Optimization;
Dimensional accuracy;
Surface quality;
Taguchi method;
Grey relational analysis

**Corresponding author:*
davorin.kramar@fs.uni-lj.si
(Kramar, D.)

Article history:
Received 28 November 2024
Revised 12 February 2025
Accepted 27 February 2025

1. Introduction

Micro-milling is extensively used for machining of inclined and free-form surfaces with very high precision, e.g., in mold manufacturing, automotive and aerospace industries, optics, biomedical industries, etc. Camara *et al.* [1] characterized the micro-milling process by the size of the cutting-edge diameter of the tool, which ranges from 1 µm and 1 000 µm. Among various applications of micro-milling, the mold making industry is one of the most important due to the rapid and accurate machining of high aspect ratio in the micro-domain [2]. The dimensional accuracy and surface roughness of the micro-parts manufactured with this micro-mechanical cutting process plays a major role in defining the quality of a die. In complex engineering environments, predicting product quality based on performance parameters represents a challenging task [3]. There are some problems associated with the micro-milling process primarily induced by excessive cutting forces and cutting tool vibrations that can deteriorate the part quality or limit the overall productivity.

These performance characteristics are highly affected by the process parameters such as depth of cut, stepover distance, cutting speed, feed rate, workpiece material type, cooling/lubrication conditions, etc. Hence, the selection of optimum control parameters is a very important step to obtain desired quality of the machined parts [4, 5].

Numerous studies have been carried out to improve different quality performance indices in micro-milling and carry out parameter's optimization. For instances, Ray *et al.* [6] conducted an experimental analysis on Zr-based bulk metallic glass to evaluate the influence of the micro-milling process parameters, such as feed per tooth, spindle speed and axial depth of cut on average line roughness, average area roughness and the dimensional accuracy of the machined micro-channels. Moreover, desirability function approach was used to determine the cutting parameters that optimizes the surface roughness and the average micro-channels width. Wojciechowski and Mrozek [7] carried out an analysis of dynamics of micro-ball end milling of hardened steel with various tool axis inclination angles. The optimization of the feed per tooth and tool inclination with cutting force components and accelerations of vibrations as objective functions were also conducted. An adaptive control optimization system to optimize feed rate and spindle speed for micro-milling (grooving) operations of AISI H13 tool steel in accordance with the estimation of tool wear state was proposed [8]. This study considers surface roughness and dimensional accuracy in terms of dimensional and form error as main aspects that define part quality. Vázquez *et al.* [9] utilized a particle swarm optimization (PSO) algorithm to identify optimal levels of micro-milling process parameters (depth per pass, axial depth of cut, spindle speed and feed) with surface roughness and geometrical and dimensional features of micro-channels fabricated on aluminium and titanium alloys as objective functions. Kuram and Ozcelik [10] utilized Taguchi based grey relational analysis (GRA) for multi-objective optimization in micro-milling of aluminium material Al 7075. The feed per tooth, spindle speed and depth of cut were studied as the process parameters, while the considered performance characteristics were surface roughness, cutting forces and tool wear. The effects of feed rate, spindle speed and depth of cut on the surface roughness, cutting forces and tool wear in micro-milling of two superalloys, namely, Ti6Al4V and Inconel 718, were investigated and optimized with Taguchi method [11]. Beruvides *et al.* [12] optimized the surface quality and machining time in micro-milling of tungsten-copper alloys using the non-dominated sorting genetic algorithm (NSGA-II). The desirability function approach has been used for optimization of the process parameters (feed, cutting speed and depth of cut) in micro milling of titanium alloy Ti-6Al-4V in order to simultaneously optimize surface roughness, tool wear and tool vibration [13]. Natarajan *et al.* [14] also used desirability function approach to obtain maximum surface quality and productivity in micro-end milling of aluminium, considering the spindle speed, feed and depth of cut as the machining parameters. The NSGA-II was employed to address multi-objective optimization problem for enhancing surface quality and dimension accuracy in micro-milling of thin-walled parts [15]. The surface quality in micro-milling of Al 2011 aluminium alloy was optimized by Cardoso and Davim [16]. Thepsonthi and Özel [17] has studied the use of PSO algorithm to optimize multiple characteristics, i.e. surface quality, tool life and burr formation, in micro-milling of Ti-6Al-4V alloy. The considered machining parameters were tool path strategy, spindle speed, feed per tooth and depth of cut. In another study [18], same authors also used PSO algorithm for optimizing the process parameters in micro-milling of titanium alloy Ti-6Al-4V for minimizing surface roughness and top burr width. Aslantas *et al.* [19] reported the use of Taguchi-based GRA in multiple parameters optimization in micro-milling for Ti-6Al-4V titanium alloy. Three cutting parameters, namely, cutting speed, feed rate and depth of cut were optimized for minimal surface roughness and burr formation. Optimization of multiple performance characteristics, such as surface quality, tool wear and tool vibration in micro-milling of AISI304 stainless steel has been conducted using a hybrid approach combining the Taguchi method-based graph theory and matrix approach and utility concept [20]. The spindle speed, depth of cut and feed rate were the cutting parameters studied in this paper. Suneesh and Sivapragash [21] identified optimal parameters for micro-milling of magnesium alloy and its alumina composites using GRA and techniques for order of preference by similarity to ideal solution. Three objectives including cutting forces, surface roughness and tool wear were considered in the optimization model, which are affected by three variables, namely spindle speed, cutting depth and feed per tooth. Miljušković and Cica [22] studied the impact of the micro-milling parameters such as depth of cut, stepover, feed and spindle speed to the mean roughness depth on a graphite electrode, followed by the differential evolution algorithm to identify the optimal machining conditions. A model based on Taguchi's signal-to-noise ratio has been used for optimization of process parameters in micro-milling of polycarbonate substrate to obtain minimum surface roughness [23]. Krimpenis *et al.* [24] employed genetic algorithm to find out optimal micro-milling process parameters with consideration of surface quality and machining time as objective function. The optimal condition of process parameters in micro milling process of hardened tool steel was found to minimize cutting forces, surface roughness, vibrations and burr formation and to maximize the material removal rate [25]. Sredanovic *et al.* [26] conducted optimization of machining parameters, including depth of cut and feed per tooth, in micro-milling of the superalloy Inconel 718 with surface roughness, cutting forces, burr formation and channel depth deviation as the optimization objectives. The slime mold sequence algorithm was suggested to solve the optimal combination of process parameters with MRR, machining cost and machining time in the CNC micro-milling process as the optimization objectives, while the machining forces, surface roughness, tool deformation and parameter uncertainty were considered as constraints [27]. Guo *et al.* [28] optimized process parameters (spindle speed, feed rate, depth of cut and tool cantilever length) for glow discharge polymer micro-milling to achieve lower cutting force and surface roughness.

Part quality is crucial to enhancing productivity, profitability and sustainability of manufacturing companies in Industry 4.0 [29]. Based on the previously mentioned literature review, the research on optimizing micro-milling parameters in terms of dimensional accuracy and surface quality has focused on the conventional metals and alloys and there is limited work available dedicated to micro-milling of graphite material in the available literature. As a result of its high thermal and chemical stability, good electrical conductivity and increasing strength with higher temperature, graphite is considered to be still the primary option for electrode materials at meso/micro scale [30]. Apart from the erosion process, the performance of the micro die-sinking electrical discharge machining process is also related to the micro-milling process of the 3D form electrodes because any potential errors are copied into a micro mold [31]. Hence, dimensional accuracy and machined surface roughness of the 3D micro die sink electrode are very important in order to attain extremely strict tolerances of a machined micro-mold. On the other hand, when milling the graphite, its inconsistent polycrystal structure undergoes localized fractures instead plastic deformation and chip formation. This process forms short fragments resulting in the formation of graphite powder, rather than chips. Thus, graphite machining has its unique characteristics dissimilar from those of metal cutting that can diminish the surface quality and the dimensional accuracy of the machined micro-features. To achieve desired level of quality characteristics, selection of optimum combination of input process parameters is crucial. However, due to complicated cutting process mechanisms linked to physical characteristics of graphite and the presence of many process factors, determination of optimal micro-milling parameters accuracy is a challenging task.

This study is based on dry micro-milling of ultrafine graphite electrodes through a set of experiments, varying four process parameters such as depth of cut, spindle speed, stepover distance and feed rate. The material used was an isostatically pressed ultrafine graphite. Results were obtained by evaluating the dimensional accuracy and arithmetic mean roughness of the part with different angles of surface inclination. Taguchi-based GRA has been employed to optimize micro-milling parameters by simultaneously minimizing dimensional deviation and minimum arithmetic mean roughness as most important indicators for high-quality manufacturing. Additionally, in the present investigation, principal component analysis was introduced to estimate actual weights of performance characteristics under optimization.

2. Research methodology and methods

The flowchart of the research approach used in the study is illustrated in Fig. 1. Firstly, the experimental plan was designed to select material, machine tool, cutting tool, micro-milling parameters and their levels and performance characteristics. The quality characteristics chosen to evaluate the processes were dimensional accuracy and arithmetic mean roughness, whereas the corresponding micro-milling parameters were depth of cut, spindle speed, stepover distance and feed rate. Experiments were performed using the Taguchi L_9 orthogonal array. The influence of micro-milling parameters on performance characteristics was determined using response surface methodology (RSM). Next, grey relational analysis (GRA) coupled with principal component analysis (PCA) has been utilized for multi-objective optimization of the machining parameters in micro ball end milling of inclined surfaces. GRA was employed to transform multiple performance characteristics into an equivalent single performance criterion, while PCA was applied to establish the corresponding weights for each performance characteristic. ANOVA was employed to analyse which of the process parameters notably affect the multiple performance characteristics.

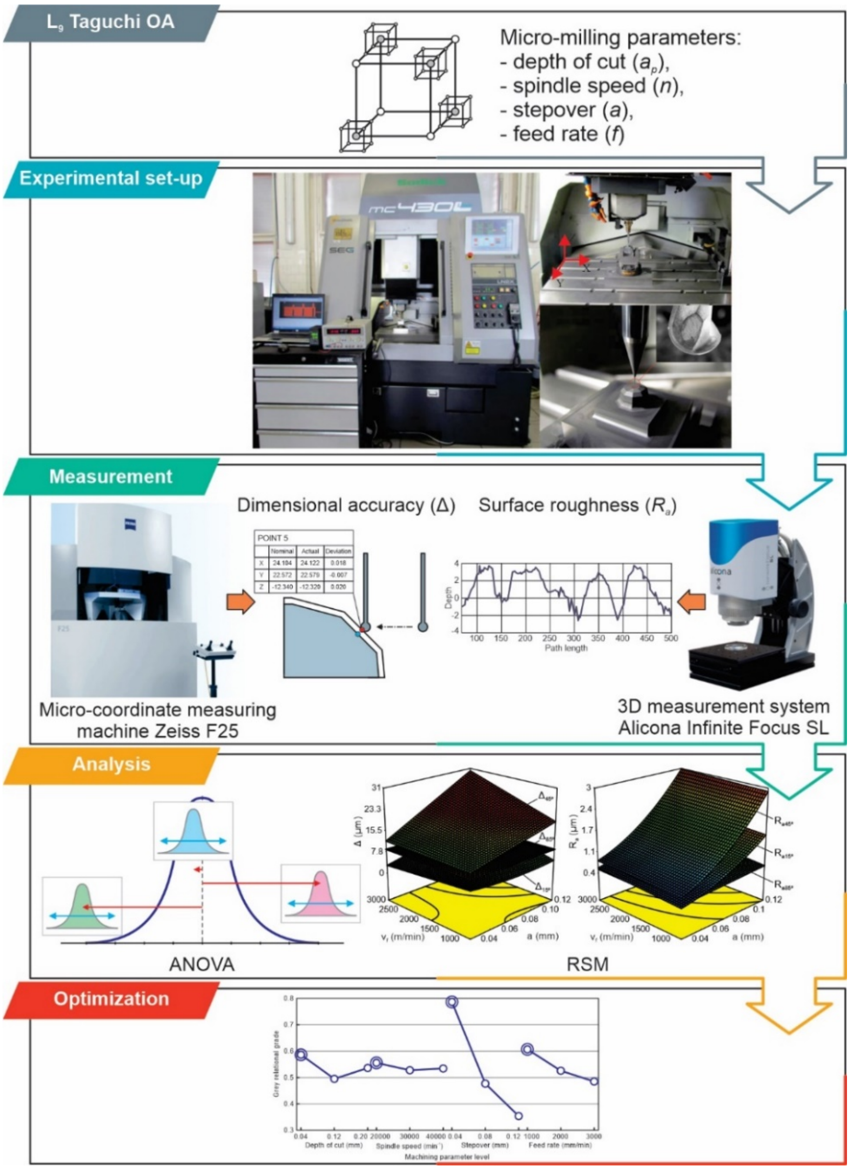


Fig. 1 Flowchart of the research methodology

3. Experimentation

The three-axis milling centre that was used to conduct the experimentation was Sodick MC430L. The maximum rotational speed is 40 000 rpm and axis travels are 420 mm, 350 mm and 200 mm for X, Y and Z axes, respectively. The feed system is fitted with linear drives and linear encoders with an absolute position accuracy of 1.5 μm. A standard heat-shrink fit HSK-E25 tool holder was employed in all experiments to minimize errors. The cutter used in the experiments was a carbide ball nose micro end-mill with a 10 μm-thick CVD diamond coating. The milling tool had a diameter d = 0.6 mm and neck length l = 6 mm.

The workpiece material tested in this study was ultrafine graphite TTK-4 (average particle size 4 μm) with the following mechanical properties: bulk density 1.78 g/cm^3, hardness 72 HSD, electrical resistivity 14 μΩ·m, tensile strength 49 MPa, flexural strength 73 MPa, compressive strength 135 MPa, Young's modulus 10.9 MPa, coefficient of thermal expansion $5\cdot10^{-6}$ K^{-1} and thermal conductivity 90 W(m·K)$^{-1}$. The sample had a rectangular prism shape with dimensions: 27 mm × 27 mm × 29 mm. To obtain three planes inclined at 15°, 45° and 85° as shown in Fig. 2, a workpiece was machined in two steps: roughing and semi-finishing to gain a determined constant depth of cut. While machining all inclined surfaces, contour operation from top to bottom in climb milling mode was carried out. All experiments were conducted without coolant to avoid contamination and air blow in a feed direction was performed throughout the experimental trials to keep the cutting zone clean.

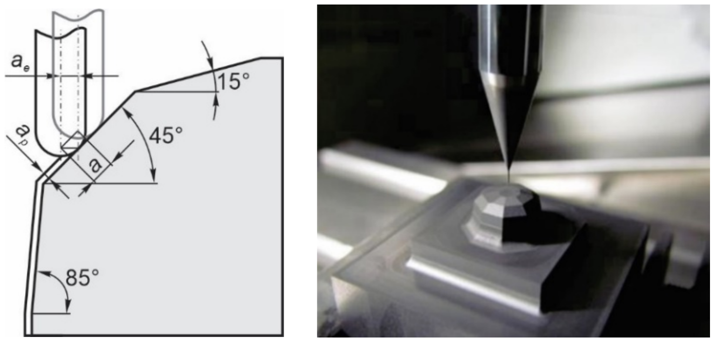


Fig. 2 The geometry of the test part

In precision engineering, coordinate measuring machines are typically used to determine the quality of dimensional and geometric part parameters [32]. Hence, dimensional accuracy was investigated by high accuracy 3D coordinate measuring machine Carl Zeiss F25, which is suitable for measuring micro-parts with linear measuring tolerance 0.5(μm) + L(mm)/666. The dimensional error of the inclined surface (Δ) was defined as the deviation between the machined and designed surface profile $\Delta = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$. Hence, the dimensional error was specified as the difference between the CAD model and the inspected part using 3D coordinate measuring machine. The dimensional error response is the average value of five measurements taken as points per surface. Although the deflection of long micro end-mills has a major impact on the dimensional accuracy of the machined part, there are also additionally sources of errors, such as positional accuracy and repeatability of the machine tool, spindle thermal expansion, radius tolerance of ball nose end-mill, diameter and tool length measuring accuracy, etc.

The arithmetic mean roughness R_a was measured by using non-destructive device InfiniteFocus Alicona because typically used stylus type surface roughness tester might damage the surface of the brittle graphite. Measurements of the arithmetic mean roughness were undertaken in the pick feed direction and the average value of surface finish at three different areas under each set of micro-milling conditions was captured as the basis for further analysis. During the measurement process, the cut-off length was selected as 0.8 mm and the sampling length as 2.5 mm.

The investigation conducted in the present study used the Taguchi method to design the experiments to reduce the number of experimentations. Moreover, this method is also effective for investigating the interactions between the control factors and the responses, as well as to find out optimal levels of cutting parameters. Four control factors including depth of cut a_p, spindle speed n, stepover distance a and feed rate v_f were considered as micro-milling parameters, while the response factors include dimensional accuracy and arithmetic mean roughness. Each control factor was divided into three levels according to the recommendations of the tool manufacturer, literature findings and trial runs. The micro-milling parameters and their levels are shown in Table 1. Then, the experimental design for four cutting parameters with three levels was arranged by Taguchi's L_9 orthogonal array as shown in Table 2.

Table 1 Design of experiments

Parameter	Unit	Levels		
		Level 1	Level 2	Level 3
Depth of cut (a_p)	mm	0.04	0.12	0.20
Spindle speed (n)	rpm	20 000	30 000	40 000
Stepover (a)	mm	0.04	0.08	0.12
Feed rate (v_f)	mm/min	1 000	2 000	3 000

4. Results and discussions

Table 2 shows the experimental results for dimensional deviation and arithmetic mean roughness. The observed values of these results along surfaces with different inclination angles can be used as a favourable indicator for determining the shape characteristics of the machined profile during micro-milling of ultrafine graphite. With the aim to analyse the effect of micro-machining parameters on measured values of the output variables, response surface models were developed. Through the backward elimination process, the final models of dimensional accuracy and arithmetic mean roughness for 15°, 45° and 85° workpiece inclination angles in a form of reduced second-order polynomials with regression coefficients are presented in Table 3. Model terms that are not significant are not included in the reduced models. However, these models contain subset of all possible effects that retain hierarchy for statistical reasons.

Table 2 Experimental results of part dimensional accuracy and arithmetic mean roughness

No.	Micro-milling parameters				Experimental data					
	a_p (mm)	n (rpm)	a (mm)	v_f (mm/min)	$\Delta_{15°}$ (µm)	$\Delta_{45°}$ (µm)	$\Delta_{85°}$ (µm)	$R_{a15°}$ (µm)	$R_{a45°}$ (µm)	$R_{a85°}$ (µm)
1	0.04	20 000	0.04	1 000	0.9	2.3	1.9	0.47	0.49	0.55
2	0.04	30 000	0.08	2 000	2.4	8.0	5.8	0.89	1.25	0.61
3	0.04	40 000	0.12	3 000	11.9	19.8	12.6	1.77	2.74	0.90
4	0.12	20 000	0.08	3 000	5.6	20.4	17.7	1.03	1.58	0.81
5	0.12	30 000	0.12	1 000	5.0	19.3	15.7	1.62	2.89	0.68
6	0.12	40 000	0.04	2 000	6.4	12.7	13.8	0.47	0.53	0.47
7	0.20	20 000	0.12	2 000	4.6	18.7	16.6	1.66	3.02	0.77
8	0.20	30 000	0.04	3 000	0.3	5.2	8.2	0.55	0.64	0.52
9	0.20	40 000	0.08	1 000	3.7	7.4	3.9	0.79	1.34	0.55

Table 3 Summary of model's coefficients

Resp.	b_0	a_p	n	a	v_f	$a_p \times n$	$n \times v_f$	$a \times v_f$	a_p^2	a^2
$\Delta_{15°}$	1.094	123.75	$1.82 \cdot 10^{-4}$	-184.6	-0.0045			0.074	-496.09	593.75
$\Delta_{45°}$	-10.43	341.67		20.833	-0.0027			0.0679	-1342.45	
$\Delta_{85°}$	-26.77	520.63	$7.79 \cdot 10^{-4}$			-0.0073			-1182.29	
$R_{a15°}$	0.3067			-3.833	$7.8 \cdot 10^{-5}$					116.667
$R_{a45°}$	0.071		$2.7 \cdot 10^{-5}$	-21.04	$5.6 \cdot 10^{-4}$		$-1.73 \cdot 10^{-8}$			313.542
$R_{a85°}$	0.586		$-3.5 \cdot 10^{-6}$	0.25	$-5 \cdot 10^{-5}$			0.00156		

Table 4 Evaluation of the models

Response	F-value	P-value	Hierarchical terms	Influential terms	R^2	R^2_{adj}	R^2_{pred}	S/N ratio
$\Delta_{15°}$	95.46	0.0295	a_p	n, a, v_f, $a \times v_f$, a_p^2, a^2	0.9998	0.9983	0.9745	87
$\Delta_{45°}$	64.43	0.0030	a_p	a, v_f, $a \times v_f$, a_p^2	0.9908	0.9745	0.9457	21.06
$\Delta_{85°}$	38.38	0.0019	a_p, n	$a_p \times n$, a_p^2	0.9746	0.9492	0.8146	15.01
$R_{a15°}$	430.83	< 0.0001	-	a, v_f, a^2	0.9961	0.9938	0.9857	48.63
$R_{a45°}$	3517.08	< 0.0001	-	n, a, v_f, $n \times v_f$, a^2	0.9998	0.9995	0.9962	141.96
$R_{a85°}$	20.22	0.0065	n	a, v_f, $a \times v_f$	0.9529	0.9057	0.7487	12.06

The statistical significances of the developed quadratic models were evaluated based on the F and P-values calculated within analysis of variance (ANOVA), as shown in Table 4.

The obtained models were regarded statistically significant when the P-values are smaller than 0.05 (95 % confidence). Moreover, models were also analysed using determination coefficient R^2, adjusted determination coefficient R^2_{adj}, predicted determination coefficient R^2_{pred} and S/N ratio.

The results show that the developed response surface models provide adequate approximation of investigated process under the given experimental domain.

The 3D surface plots of dimensional accuracy are presented in Fig. 3. The highest values of dimensional error were observed at 45° workpiece inclination angle, whereas the best dimensional accuracy is evident at the 15° inclined plane for all considered parameters. Tool deflection caused by cutting forces is considered as the main factor that influences machining error [33]. As the diameter the of micro ball-end milling tool is extremely small, the stiffness is most sensitive to influence by cutting force than any other parameter and consequently bending deformation. The cutting forces generated while micro-milling of inclined surface cause tool deflection that leads to the form error of part surface. The cutting force components are affected by the workpiece's inclination angle. Axial force decreases and radial force increases in magnitude as the inclination angle increases. The change in the workpiece inclination angle has a significant effect on the tool deflection as a result of the lower stiffness in the radial as compared to the axial direction which is attributed to the lower stiffness. Consequently, better dimensional accuracy can be achieved with lower workpiece inclination angle values. The variation of the dimensional accuracy with depth of cut and spindle speed indicate that lower values of both parameters lead to smaller dimensional errors (Fig 3a). As depth of cut and spindle speed increases, that results increase in the MRR, that has a positive correlation with cutting forces. The experimental results prove that the dimensional error of the machined surface is considerably influenced by the stepover, as shown in Fig. 3b. The results indicate that smaller values of stepover distance will lead to a significant increase in the dimensional accuracy for all workpiece inclination angle values. An increase in stepover during micro-milling operation leads to an increase in the metal removal volume and therefore to an increase of cutting forces. Higher cutting forces cause a larger tool deflection which results in higher dimensional errors. Moreover, a similar trend for feed rate was observed as in stepover distance, where increasing the values of feed rate results in higher dimensional deviations. The reason being, increased feed rate value leads to large chip sizes and hence the growth of the cutting forces in micro-milling operation. Moreover, an increase in the feed rate also results in an increase of self-excited vibration (chatter). Subsequently, an increase in cutting forces and vibrations results in large values of cutting tool deflections which lead to geometric errors on the machined part.

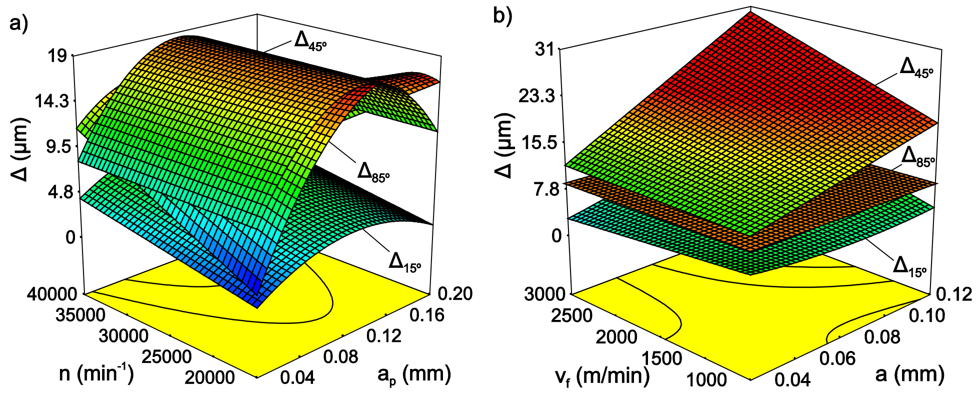


Fig. 3 Effect of micro-milling parameters on dimensional accuracy

Fig. 4 show the 3D surface plots of the arithmetic mean roughness. From this figure, it is seen that the highest arithmetic mean roughness was observed for 45° workpiece inclination angle, while the best machined surface finish is noted for 85° inclination angle. This was because during machining remarkably steep surfaces (almost vertical), the cylindrical segment of the ball end mill is mostly in contact with the machined surface causing generation of linear cusps with lower profile height as compared to spherical shaped cusps. As viewed in surface response in Fig. 4a, the highest surface quality is obtained with the combination of the highest spindle speed and the lowest feed rate. Increased spindle speed results in higher tooth passing frequencies and shorter plane area/reduction in chip thickness, lowering surface roughness. The increase in feed rate increases the heat generation and vibration due to increase in MRR, leading to higher surface roughness. Fig. 4b shows the interaction effect between the stepover and the feed rate on the surface quality. The stepover distance is the most significant factor associated with the arithmetic mean roughness. The surface finish significantly improved with decrease in stepover distance. This variation is identically changed for all workpiece inclination angle values. This can be attributed to the fact that in the ball end milling process the stepover defines the overall peripheral area of the cutting tool which is in contact with the workpieces surface. The increase in stepover value increases overlap between cutting paths and it produces higher cusps height of the machined surface resulted in a deterioration of surface quality. Hence, smaller values of stepover distance must be selected to achieve the better arithmetic mean roughness. It is seen that the arithmetic mean roughness decreases with the decrease in the feed rate. This phenomenon can be explained by the higher cutting forces and the heat generation due to the larger cutting area at high feed rate. Besides that, increase in feed rate also increases the chatter resulting in poor surface finish.

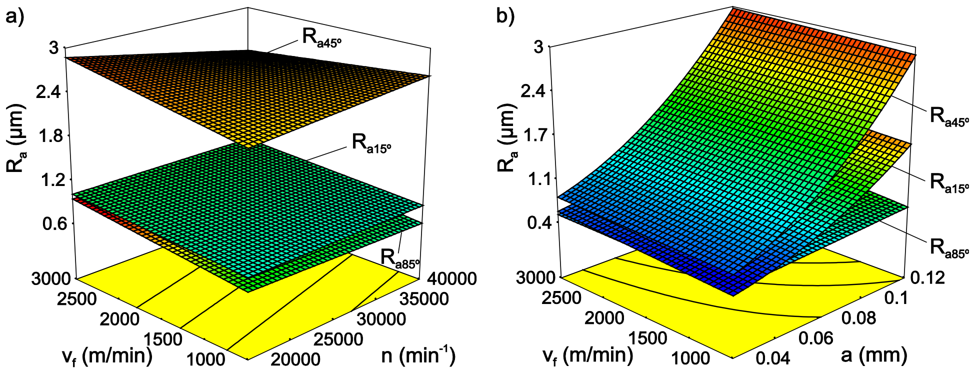


Fig. 4 Effect of micro-milling parameters on the arithmetic mean roughness

5. Multi-objective optimization of micro-milling process by Taguchi based grey relational analysis

As depicted in the previous section, dimensional accuracy of part produced by micro-milling operation as well as surface quality varies significantly with the changes in the machining parameters. Presented multi-objective optimization method aimed to obtain the optimum micro-milling parameters to minimize both dimensional errors and the arithmetic mean roughness. The performance characteristics, i.e. dimensional accuracy and arithmetic mean roughness for 15°, 45° and 85° workpiece inclination angles, obtained from the experimental results, are firstly converted into S/N ratio. The S/N ratios matching to each of the studied single quality characteristics are normalized, because larger values of the normalized results indicate better performance. Thereafter, the grey relational coefficients were calculated.

Frequently, in order to calculate the GRG, the equal weight factors of each performance characteristic were selected for simplicity. However, this approach may not be an appropriate, due to fact that the importance of various quality characteristics is different in real engineering problems. Hence, PCA was employed to determine the appropriate weight of each performance characteristic. The grey relational coefficients of each performance characteristics have been used to determine the correlation coefficient matrix and calculate the corresponding eigenvalues. The variance contribution for the first principal component characterizing the six performance characteristics is as high as 68.2 %. Accordingly, the squares of corresponding eigenvectors were chosen as the weighting factors of the associated performance characteristic and the coefficients w_1, w_2, w_3, w_4, w_5 and w_6 are consequently set as 0.0952, 0.1829, 0.1358, 0.2198, 0.2241 and 0.1423, respectively. Finally, the grey relational grades were calculated by multiplying grey relational coefficients with their corresponding weight of performance characteristics.

Table 5 The calculated grey relation coefficients and grey relational grades for six different machining responses

No.	Grey relational coefficient						Grey relational grade	Grey order
	$\Delta_{15°}$ (μm)	$\Delta_{45°}$ (μm)	$\Delta_{85°}$ (μm)	$Ra_{15°}$ (μm)	$Ra_{45°}$ (μm)	$Ra_{85°}$ (μm)		
1	0.6262	1.0000	1.0000	1.0000	1.0000	0.6739	0.9181	1
2	0.4695	0.4668	0.5000	0.5094	0.4926	0.5547	0.4993	5
3	0.3333	0.3364	0.3710	0.3333	0.3457	0.3333	0.3418	9
4	0.3860	0.3333	0.3333	0.4580	0.4371	0.3737	0.3948	6
5	0.3954	0.3391	0.3457	0.3489	0.3388	0.4679	0.3658	7
6	0.3755	0.3898	0.3601	1.0000	0.9206	1.0000	0.7244	2
7	0.4027	0.3424	0.3398	0.3444	0.3333	0.3969	0.3540	8
8	1.0000	0.5722	0.4328	0.8084	0.7730	0.7626	0.7181	3
9	0.4228	0.4829	0.6081	0.5608	0.4748	0.6739	0.5367	4

Table 5 shows the calculated grey relational coefficients and grey relational grades based for each experiment using the Taguchi L_9 orthogonal array. From Table 5, it has been noted that experiment No. 1 has the highest value of grey relational grade as 0.9180, whereas the lowest value was found for experiment No. 3 as 0.3418.

Furthermore, the means of the weighted grey relational grade of each micro-machining parameter have been computed and listed in Table 6 and depicted in Fig. 5. From the analysis of the response table and main effect plot for weighted grey relational grade, the optimal level setting of micro-machining parameters is as follows a_p = 0.04 mm, n = 20 000 min^{-1}, a = 0.04 mm and v_f = 1 000 mm/min. Thus, the optimal combination of micro-milling parameters for minimum dimensional deviation and minimum arithmetic mean roughness under the given experimental design was obtained when they are at their minimal level. This optimal process parameters setting that optimize the considered multiple objective function corresponds to experiment No. 1 shown in Table 2. Nevertheless, the relative significance among the micro-milling process parameters for optimized the quality indicators needs to be further analysed and understood to obtain the best parametric combination more clearly. Apart from analysis of the means accomplished for the obtained grey relational grade, in Table 6 is also listed the rank of the micro-milling parameter affecting the grey relational grade. The grey relational grade for each control factor is ranked according to the difference between its maximum and minimum values. This difference can be also defined as the effect contribution of machining parameters. The response table indicates that stepover distance has the maximum level difference value of grey relational grade. Hence, this is the most influential factor affecting the overall characteristic. The second most significant factor is feed rate, followed by the depth of cut and spindle speed that has the least prominent effect on the multi-performance characteristic. These values are depicted graphically in Fig. 5. The graph indicates that higher levels of depth of cut, spindle speed, stepover distance and feed rate have negative effect on the weighted grey relational grade, that is, dimensional inaccuracy and arithmetic mean roughness increase with increase in value of these micro-milling parameters.

An analysis of variance (ANOVA) was conducted to determine the influence of each machining parameter on the multi-performance characteristic. Since the degrees of freedom (DOF) for residual error was zero, the test for significance is not possible. Consequently, ANOVA pooling was performed. Pooling is the process of merging the influence of the insignificant factors with the error term to create a new error term that can be tested further. Normally, this occurs because selected orthogonal array L_9 with four parameters varied through three levels does not provide enough data. In general, pooling process start with the factor that has the least influence. In present paper, spindle speed was found insignificant (pooled). From the pooled ANOVA table (Table 7), it is obvious that the stepover distance is the most significant factor that contributes towards the overall performance characteristic, since it contributes to the highest percentage of variation of 89 %. This is followed by feed rate and depth of cut which contribute 6.9 % and 3.7 %, respectively. The percentage of error was considerably low at 0.4 %.

Table 6 The response table for weighted grey relational grade

Control parameter	Grey relational grade			Max-min	Rank
	Level 1	Level 2	Level 3		
Depth of cut (a_p)	0.5864	0.4950	0.5363	0.0914	3
Spindle speed (n)	0.5556	0.5277	0.5343	0.0279	4
Stepover (a)	0.7869	0.4769	0.3539	0.4330	1
Feed rate (v_f)	0.6069	0.5259	0.4849	0.1220	2
Total mean of the grey relational grade: 0.5393					

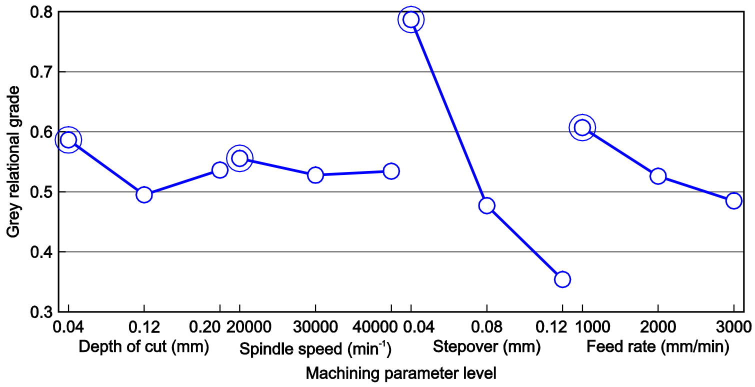


Fig. 5 Main effect plot for weighted grey relational grade

Table 7 Results of the pooled ANOVA for grey relational grade

Source	Sum of squares	DOF	Mean square	*F*-value	*P*-value	PC (%)
a_p	0.01257	2	0.00629	9.85	0.0922	3.7
a	0.29869	2	0.14935	233.97	0.0043	89.0
v_f	0.02311	2	0.01156	18.1	0.0523	6.9
Residual	0.00128	2	0.00064			0.4
Total	0.33565	8				100

5. Conclusion

In this study, experimental investigation and multi-objective optimization of the dry micro-milling process while machining of ultrafine graphite TTK 4 has been carried out. Response surface methodology was used to investigate effects of depth of cut, spindle speed, stepover distance and feed rate on dimensional deviation and arithmetic mean roughness along surfaces with three angles of inclination: 15°, 45° and 85°, whereas the grey relational analysis coupled with principal component analysis was employed to optimize the process parameters. Based on the experimental findings of this study, the following conclusions can be drawn:

- Workpiece inclination angle was discovered to have a large influence on dimensional deviation and arithmetic mean roughness in micro ball-end milling process. Best dimensional accuracy was observed at the smallest angle of inclination that is 15°, followed by 85° and 45° workpiece inclination angles. On the other hand, minimum arithmetic mean roughness was observed for the 85° inclination angle of workpiece, while maximum value of this parameter was achieved at the inclination angle of 45°.
- Taguchi based grey relational analysis greatly simplifies the optimization of the multi-response problems due to conversion of the multiple performance characteristics into single performance measure. Moreover, principal component analysis is a well-suited technique for obtaining the corresponding weighting values of each performance characteristics.
- Optimization procedure revealed that the optimum micro-milling conditions for minimum dimensional deviation and minimum arithmetic mean roughness were at a low level of depth of cut, spindle speed, stepover distance and feed rate, i.e. a_p = 0.04 mm, n = 20 000 min^{-1}, a = 0.04 mm and v_f = 1 000 mm/min.

- The results of ANOVA showed that the stepover distance is the most influential factor among the four micro-milling parameters used on the multi-performance characteristics contributing by 89 %, while feed rate and depth of cut contribute 6.9 % and 3.7 %, respectively.

The analysis and subsequent optimization of micro-milling of ultrafine graphite electrodes with high aspect ratio and different angles of surface inclination was performed in this research. These results are expected to be valuable for micro-milling of other brittle materials, for example, silicon, glass, etc. In future works, the experimental area can potentially be expanded and more response variables such as tool wear, work materials, etc. could be included. Finally, it will be interesting to examine machining of more complex surfaces.

References

[1] Câmara, M.A., Campos Rubio, J.C., Abrão, A.M., Davim, J.P. (2012). State of the art on micro-milling of materials, a review, *Journal of Materials Science & Technology*, Vol. 28, No. 8, 673-685, doi: 10.1016/S1005-0302(12)60115-7.

[2] O'Toole, L., Kang, C.-W., Fang, F.-Z. (2021). Precision micro-milling process: State of the art, *Advances in Manufacturing*, Vol. 9, 173-205, doi: 10.1007/s40436-020-00323-0.

[3] Pang, J.H., Zhao, H., Qin, F.F., Xue, X.B., Yuan, K.Y. (2019). A new approach for product quality prediction of complex equipment by grey system theory: A case study of cutting tools for CNC machine tool, *Advances in Production Engineering & Management*, Vol. 14, No. 4, 461-471, doi: 10.14743/apem2019.4.341.

[4] Kramar, D., Cica, Dj. (2021). Modeling and optimization of finish diamond turning of spherical surfaces based on response surface methodology and cuckoo search algorithm, *Advances in Production Engineering & Management*, Vol. 16, No. 3, 326-334, doi: 10.14743/apem2021.3.403.

[5] Vukelic, D., Milosevic, A., Ivanov, V., Kocovic, V., Santosi, Z., Sokac, M., Simunovic, G. (2024). Modelling and optimization of dimensional accuracy and surface roughness in dry turning of Inconel 625 alloy, *Advances in Production Engineering & Management*, Vol. 19, No. 3, 371-385, doi: 10.14743/apem2024.3.513.

[6] Ray, D., Puri, A.B., Hanumaiah, N. (2020). Experimental analysis on the quality aspects of micro-channels in mechanical micro milling of Zr-based bulk metallic glass, *Measurement*, Vol. 158, Article No. 107622, doi: 10.1016/j.measurement.2020.107622.

[7] Wojciechowski, S, Mrozek, K. (2017). Mechanical and technological aspects of micro ball end milling with various tool inclinations, *International Journal of Mechanical Sciences*, Vol. 134, 424-435, doi: 10.1016/j.ijmecsci.2017.10.032.

[8] Coppel, R., Abellan-Nebot, J.V., Siller, H.R., Rodriguez, C.A., Guedea, F. (2016). Adaptive control optimization in micro-milling of hardened steels-evaluation of optimization approaches, *The International Journal of Advanced Manufacturing Technology*, Vol. 84, 2219-2238, doi: 10.1007/s00170-015-7807-6.

[9] Vázquez, E., Ciurana, J., Rodríguez, C.A., Thepsonthi, T., Özel, T. (2011). Swarm intelligent selection and optimization of machining system parameters for microchannel fabrication in medical devices, *Materials and Manufacturing Processes*, Vol. 26, No. 3, 403-414, doi: 10.1080/10426914.2010.520792.

[10] Kuram, E., Ozcelik, B., (2013). Multi-objective optimization using Taguchi based grey relational analysis for micro-milling of Al 7075 material with ball nose end mill, *Measurement*, Vol. 46, No. 6, 1849-1864, doi: 10.1016/j.measurement.2013.02.002.

[11] Kuram, E., Ozcelik, B. (2015). Optimization of machining parameters during micro-milling of Ti6Al4V titanium alloy and Inconel 718 materials using Taguchi method, *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, Vol. 231, No. 2, 228-242, doi: 10.1177/0954405415572662.

[12] Beruvides, G., Castaño, F., Quiza, R., Haber, R.E. (2016). Surface roughness modeling and optimization of tungsten-copper alloys in micro-milling processes, *Measurement*, Vol. 86, 246-252, doi: 10.1016/j.measurement.2016.03.002.

[13] Venkata Rao, K. (2019). A study on performance characteristics and multi response optimization of process parameters to maximize performance of micro milling for Ti-6Al-4V, *Journal of Alloys and Compounds*, Vol. 781, 773-782, doi: 10.1016/j.jallcom.2018.12.105.

[14] Natarajan, U., Periyanan, P.R., Yang S.H. (2011). Multiple response optimization for micro-end milling process using response surface methodology, *The International Journal of Advanced Manufacturing Technology*, Vol. 56, 177-185, doi: 10.1007/s00170-011-3156-2.

[15] Wang, P., Bai, Q., Cheng, K., Zhao, L., Ding, H. (2022). Optimization of the process parameters for micro-milling thin-walled micro-parts using advanced algorithms, *The International Journal of Advanced Manufacturing Technology*, Vol. 121, 6255-6269, doi: 10.1007/s00170-022-09729-5.

[16] Cardoso, P., Davim, J.P. (2010). Optimization of surface roughness in micromilling, *Materials and Manufacturing Processes*, Vol. 25, No. 10, 1115-1119, doi: 10.1080/10426914.2010.481002.

[17] Thepsonthi, T., Özel, T. (2014). An integrated toolpath and process parameter optimization for high-performance micro-milling process of Ti-6Al-4V titanium alloy, *The International Journal of Advanced Manufacturing Technology*, Vol. 75, 57-75, doi: 10.1007/s00170-014-6102-2.

[18] Thepsonthi, T., Özel, T. (2012). Multi-objective process optimization for micro-end milling of Ti-6Al-4V titanium alloy, *The International Journal of Advanced Manufacturing Technology*, Vol. 63, 903-914, doi: 10.1007/s00170-012-3980-z.

[19] Aslantas, K., Ekici, E., Çiçek, A. (2018). Optimization of process parameters for micro milling of Ti-6Al-4V alloy using Taguchi-based gray relational analysis, *Measurement*, Vol. 128, 419-427, doi: 10.1016/j.measurement.2018.06.066.

[20] Brahmeswara Rao, D., Venkata Rao, K., Gopala Krishna, A. (2018). A hybrid approach to multi response optimization of micro milling process parameters using Taguchi method based graph theory and matrix approach (GTMA) and utility concept, *Measurement*, Vol. 120, 43-51, doi: 10.1016/j.measurement.2018.02.005.

[21] Suneesh, E., Sivapragash, M. (2021). Multi-response optimisation of micro-milling performance while machining a novel magnesium alloy and its alumina composites, *Measurement*, Vol. 168, Article No. 108345, doi: 10.1016/j.measurement.2020.108345.

[22] Miljušković, G. Cica, Dj. (2021). Investigation, modeling and optimization of surface roughness in micro-milling of graphite electrodes, *The International Journal of Advanced Manufacturing Technology*, Vol. 117, 579-590, doi: 10.1007/s00170-021-07762-4.

[23] Chen, P.-C., Pan, C.-W., Lee, W.-C., Li, K.-M. (2014). Optimization of micromilling microchannels on a polycarbonate substrate, *International Journal of Precision Engineering and Manufacturing*, Vol. 15, 149-154, doi: 10.1007/s12541-013-0318-1.

[24] Krimpenis, A.A., Fountas, N.A., Ntalianis, I., Vaxevanidis, N.M. (2014). CNC micromilling properties and optimization using genetic algorithms, *The International Journal of Advanced Manufacturing Technology*, Vol. 70, 157-171, doi: 10.1007/s00170-013-5248-7.

[25] Balázs, B.Z., Takács, M. (2020). Experimental investigation and optimisation of the micro milling process of hardened hot-work tool steel, *The International Journal of Advanced Manufacturing Technology*, Vol. 106, 5289-5305, doi: 10.1007/s00170-020-04991-x.

[26] Sredanovic, B., Cica, Dj., Borojevic, S., Tesic, S., Kramar, D. (2024). Optimization of superalloy Inconel 718 micro-milling process by combined Taguchi and multi-criteria decision making method, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, Vol. 46, Article No. 423, doi: 10.1007/s40430-024-04996-7.

[27] Ding, P., Huang, X., Zhang, X., Li, Y., Wang, C. (2022). Reliability optimization of micro-milling cutting parameters using slime mould sequence algorithm, *Simulation Modelling Practice and Theory*, Vol. 119, Article No. 102575, doi: 10.1016/j.simpat.2022.102575.

[28] Guo, R., Chen, M., Wang, G., Zhou, X. (2022). Milling force prediction and optimization of process parameters in micro-milling of glow discharge polymer, *The International Journal of Advanced Manufacturing Technology*, Vol. 122, 1293-1310, doi: 10.1007/s00170-022-09951-1.

[29] BouAbid, H., Dhouib, K., Gharbi, A. (2024). Integrated production and maintenance policy for manufacturing systems prone to products' quality degradation, *Advances in Production Engineering & Management*, Vol. 19, No. 4, 512-526, doi: 10.14743/apem2024.4.521.

[30] Huo, D., Lin, C., Dalgarno, K. (2014). An experimental investigation on micro machining of fine-grained graphite, *The International Journal of Advanced Manufacturing Technology*, Vol. 72, 943-953, doi: 10.1007/s00170-014-5730-x.

[31] Miljušković, G., Krajnik, P., Kopač, J. (2015). Analysis of tool deflection in micro milling of graphite electrodes, *The International Journal of Advanced Manufacturing Technology*, Vol. 76, 209-217, doi: 10.1007/s00170-013-5536-2.

[32] Štrbac, B., Ranisavljev, M., Orošnjak, M., Havrlišan, S., Dudić, B., Savković, B. (2024). Unsupervised machine learning application in the selection of measurement strategy on coordinate measuring machine, *Advances in Production Engineering & Management*, Vol. 19, No. 2, 209-222, doi: 10.14743/apem2024.2.502.

[33] Cica, Dj., Sredanovic, B., Miljuskovic G. (2022). Experimental investigation of tool deflection in micro-milling of fine-grained graphite, *The International Journal of Advanced Manufacturing Technology*, Vol. 123, 161-168, doi: 10.1007/s00170-022-10185-4.

APEM journal

Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 87–98
https://doi.org/10.14743/apem2025.1.529

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Enhancing aerospace products quality with ISOMAP key factor identification

Shen, D.Y.[a,b], Liu, N.Z.[a,*], Liu, W.[b], Wang, Z.[b]

[a]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, P.R. China
[b]Beijing System Design Institute of the Electro-mechanic Engineering, Beijing, P.R. China

A B S T R A C T

The nonlinear and high-dimensional nature of the impact factors affecting the production quality of aerospace products represents a major difficulty for the quality control in the aerospace industry. To obtain the impact factors affecting the product quality, it is plausible to perform dimensionality reduction on acquired samples before further manipulation. In this work, the isometric feature mapping (ISOMAP) algorithm of stream learning is employed to perform nonlinear dimensionality reduction on aerospace data. This enables a calculation of the correlation coefficients between the principal components after dimensionality reduction and the original factors, the classification of the correspondence, and the ranking of the principal components according to their degree of influence. The experimental results show that the algorithm is able to carry out correlation analysis of 17 factors affecting the production quality of aerospace products, and analyze the 13 main factors affecting the production quality of aerospace products, and the degree of influence, in descending order, are the rationality of measurement methods, the rationality of test point design, tool wear, equipment normalization rate, the degree of equipment aging, the rationality of program design, the degree of material defects, the rationality of process route design, the rationality of tooling design, the technical level of personnel, the level of personnel experience, the personnel work status, and operational standardization. The ISOMAP algorithm was used to reduce the dimensionality of these 13 factors to form and rank the six main influence components, thus eliminating redundant factors, highlighting main influence features and extracting the intrinsic relation in data. The data analysis conclusions can facilitate a prevention of potential quality issues in aerospace production. To ensure the enhancement of the quality of aerospace product production, it is recommended that standard automated measurement methods be employed wherever feasible. Additionally, it is recommended that the regular maintenance of machining tools and equipment be strengthened to ensure that the machining tools and equipment are in perfect condition.

A R T I C L E   I N F O

Keywords:
Aerospace products;
Production quality;
ISOMAP;
Nonlinear dimensionality reduction;
5M1E framework;
Correlation analysis;
Quality control

**Corresponding author:*
liunz_tx@163.com
(Liu, N.Z.)

Article history:
Received 28 October 2024
Revised 8 April 2025
Accepted 21 April 2025

1. Introduction

The development of aerospace products is contingent upon the attainment of exceptionally high reliability standards. Consequently, the supervision and prevention of the production quality of aerospace products represents a pivotal aspect of the aerospace product development process. The production of aerospace products is susceptible to the potential for the creation of substandard items if the procedures and equipment utilized are not sufficiently standardized. With the rapid development of manufacturing intelligence technology, big data, artificial intelligence and other new technologies are deeply integrated with manufacturing, becoming an important support for the transformation and upgrading of manufacturers [1-3]. Currently, research on applying big data and artificial intelligence technologies to manufacturing focuses on optimizing the power consumption of integrated systems for flexible manufacturing [4], increasing production capacity [5], optimizing manufacturing schedules [6, 7], analyzing nonconforming production [8], evaluating production efficiency [9] and classifying faults [10]. Mature methods have been developed in these areas. Aerospace product manufacturers have implemented digital production through the implementation of a MES system, thereby accumulating substantial data regarding the production process and forming industrial big data [11, 12]. However, the question of how to utilize this data to identify the main factors affecting the production quality of aerospace products remains a significant challenge that requires immediate attention.

The main factors affecting the production quality of aerospace products are typically nonlinear and high-dimensional data, which pertain to all aspects of the production process. The use of high-dimensional data presents a number of evident challenges when undertaking decision-making analysis. Firstly, the processing of high-dimensional data is inherently challenging due to its multidimensional nature, which often results in inefficiencies in computational operations. Secondly, the presence of redundancy between factors makes it challenging to ascertain the contribution rate of each factor. Furthermore, it is difficult to identify the specific affecting factors that truly determine the quality of the production.

In the context of high-dimensional data processing, the fundamental dimensionality reduction algorithms can be classified into two categories: linear and nonlinear ones. Linear dimensionality reduction algorithms mainly include Principal Component Analysis (PCA) algorithm [13] and Multi-Dimensional Scaling (MDS) algorithm [14]. PCA algorithm projects N-dimensional features to a K-dimensional orthogonal space through a linear transformation, thereby maximizing the variance of the projected data. The obtained K orthogonal components are designated as the principal component. Currently, PCA has been widely used in network abnormal traffic detection [15], port throughput forecasting [16], image classification [17], predictions in precision agriculture [18], coin classification [19], etc. The core idea of the MDS algorithm is to utilize the distance matrix to illustrate the degree of similarity or the correlation between the data points. In contrast to the PCA algorithm, the MDS algorithm preserves the distance relationship between the original data points throughout the dimensionality reduction process. While, the PCA algorithm prioritizes the preservation of the primary trends within the data. With the advancement of computer technology, artificial intelligence technology, bioinformatics technology, and other multidisciplinary application technologies, an increasing number of high-dimensional data sets exhibit nonlinear structural characteristics. Consequently, linear dimensionality reduction algorithms have to be extended to effectively restore the low-dimensional structure in nonlinear data.

To address the challenge of nonlinear dimensionality reduction within the linear dimensionality reduction framework, several nonlinear dimensionality reduction methods have been proposed. For instance, Shawe-Taylor *et al*. employed nonlinear dimensionality reduction algorithms based on kernel methods to conduct research on nonlinear high-dimensional data dimensionality reduction [20]. Schölkopf *et al*. proposed the kernel principal component analysis (KPCA) [21], which has since become a widely used technique in fields such as face recognition [22], speech recognition [23] and novelty detection [24]. Maaten *et al*. proposed the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [25]. The t-SNE algorithm is primarily employed for the analysis of local data structures, with a focus on the extraction of local clusters. This capability is particularly advantageous for the visualization of high-dimensional data sets comprising multiple streams of varying types (e.g., the MNIST dataset). However, it is important to note its limitation in preserving the global structure of the data. The t-SNE algorithm finds primary application in fields such as bioinformatics, image processing, and related domains. McInnes *et al*. proposed the Uniform Manifold Approximation and Projection (UMAP) algorithm [26], which was developed to address the limitations of t-SNE. t-SNE has been shown to lack scalability for large data sets, does not support model persistence, and does not preserve the global structure. The UMAP algorithm has been demonstrated to be well-suited for large-scale data downscaling and visualization applications. Tenenbaum *et al*. put forth the concept of ISOMAP (Isometric Mapping) [27], which integrates PCA and MDS. Thereafter, the geodesic distance matrix has been used as the input to the MDS algorithm, leading to an improved performance.

Presently, the ISOMAP algorithm has been employed in a variety of applications, including generating training parameters for ROMs [28], fault diagnosis [29], equipment condition analysis [30], EEG classification analysis [31], groundwater systems analysis [32], and other fields. The findings from these applications have been encouraging, suggesting the potential of the ISOMAP algorithm in a variety of domains. Despite extensive research efforts have been paid on various dimensionality reduction algorithms, a notable gap persists in applying these algorithms to the production quality analysis of aerospace products.

The analysis of the main factors affecting the production quality of aerospace products is a typical small sample data analysis scenario, and the factors affecting the production quality of aerospace products have strong non-linear coupling relationships. This paper presents an analysis of the production features of aerospace products and the various types of quality-affecting factors. Compared with algorithms such as t-SNE and UMAP, the ISOMAP algorithm is more suitable for the small sample non-linear data dimension reduction scenario in the analysis of the main factors affecting the production quality of aerospace products. To achieve this, the ISOMAP algorithm is employed to carry out nonlinear dimensionality reduction, enabling a calculation of correlation coefficients between the principal component after dimensionality reduction and the original factors, a clarification of the correspondence, and a ranking of the principal component according to the degree of influence. This approach allows for the identification of the main factors that most affect the production quality of aerospace products. The algorithm is capable of effectively eliminating redundant factor interference, highlighting the main affecting features, and obtaining the inner law of the data. This can assist in addressing relevant issues in a timely manner and preventing potential quality issues in production.

2. Algorithm design

In this paper, the ISOMAP algorithm is employed for the purpose of analyzing the production quality data of aerospace products through the use of dimensionality reduction, with the objective of identifying the main affecting factors. The design of the algorithm is presented in Fig. 1.

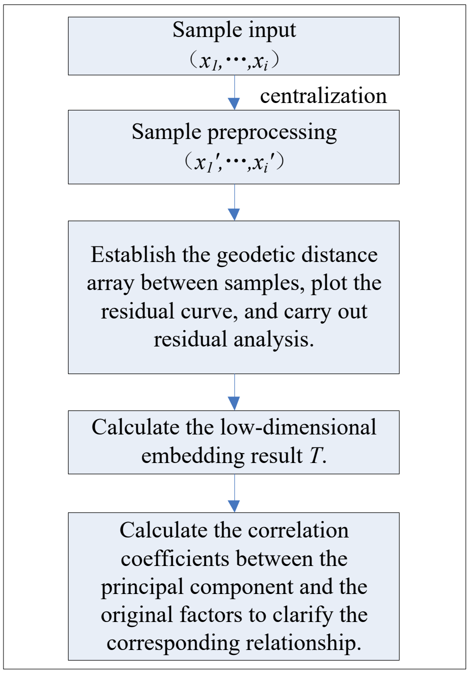


Fig. 1 Algorithm for production quality analysis of aerospace products

The ISOMAP-based algorithm for production quality analysis of aerospace products comprises the following steps: sample preprocessing, residual analysis, calculation of low-dimensional embedding results, calculation of correlation coefficients between the principal component and the original factors, and clarification of the corresponding relationship.

Step 1: Sample preprocessing

In accordance with Eq. 1, the input samples ($x_1, \ldots, x_i$) are normalized so that the transformed data are mapped between [0,1], thereby eliminating the influence of disparate factors due to the discrepancy in magnitude and size of the values. The transformed data ($x_1', \ldots, x_i'$) are then obtained.

$$x_i' = \frac{x_i - x_{imin}}{x_{imax} - x_{imin}} \tag{1}$$

Step 2: Establish the geodetic distance array between samples, plot the residual curve, and carry out residual analysis

The residual curve is plotted in accordance with Eq. 2, wherein D_g represents the geodetic distance array, D_Y denotes the Euclidean distance array, and R signifies the correlation coefficient ($R = \rho_{x,y}$). Based on the residual curve, the sample eigen-dimension d is determined. Eq. 3 illustrates the calculation of the correlation coefficient, wherein: ρ is the correlation coefficient; E is the mathematical expectation; X is the composite factor vector; and Y is the original factor vector.

$$e_d = 1 - R^2(D_g, D_y) \tag{2}$$

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{3}$$

Step 3: The results of the low-dimensional embedding are calculated

In accordance with Eq. 4, the low-dimensional embedding result T, that is to say the eigen samples ($z_1, \ldots, z_d$), is computed. Where: ($\lambda_1, \ldots, \lambda_d$) represents the largest d eigenvalue of D_g, and its corresponding eigenvector is ($u_1, \ldots, u_d$). The matrix U is defined as ($u_1, \ldots, u_d$). As the eigenvalue increases, the importance of each factor increases.

$$T = diag\left(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2}\right) U^T \tag{4}$$

Step 4: The correlation coefficients between the principal component and the original factors are calculated

The correlation coefficients between the principal component and the original factors are calculated using Eq. 3. By comparing the size of the correlation coefficients, the most relevant original factors to the principal component after dimensionality reduction are identified.

3. Example analysis

An analysis of the production process for a given product is conducted within an aerospace company. The experiment obtains the on-site processing information of this product in recent years from the MES system and takes 500 samples, including qualified and unqualified samples of 250 cases for verification. The experiment employs the PyCharm 2020 runtime environment.

3.1 Sample selection and pre-processing

The production process of aerospace products is influenced by a multitude of factors, including the interactions between men, machines, materials, methods, measurements, and the environment (5M1E). Each of these categories contains a multitude of affecting factors that collectively determine the quality of the product and its qualification status. These factors are causal regarding quality outcomes.

By combing the historical records of the production process of aerospace products in an aerospace company, we have identified the factors that influence the quality of aerospace product production around the 5M1E and determined the threshold value of each factor. This initial phase of the study represents the fundamental research component, providing the foundation for subsequent analysis of the main factors affecting the production quality of aerospace products. The factors affecting the production quality of aerospace products are presented in Fig. 2.

In terms of man factors, the following variables have been identified as affecting factors: the technical level of personnel, the level of personnel experience and the personnel work status. The technical level of personnel reflects the technical level of production personnel, encompassing primary, intermediate, and senior levels. The level of personnel experience is defined as the work experience of the production personnel involved in manufacturing and processing. According to the number of years of experience, it can be divided into three levels: level 1 (0-2 years of experience), level 2 (3-5 years of experience), and level 3 (more than 5 years of experience). Personnel work status refers to the personal status of production personnel while performing production work, reflects the degree of personal fatigue, and is closely related to individual continuous working time. The personnel work status is divided into two categories: "good" and "bad." If the personnel work continuously for more than eight hours or for more than five days, the status is designated as "bad." Conversely, if the personnel work continuously for less than eight hours or for less than five days, the status is designated as "good."

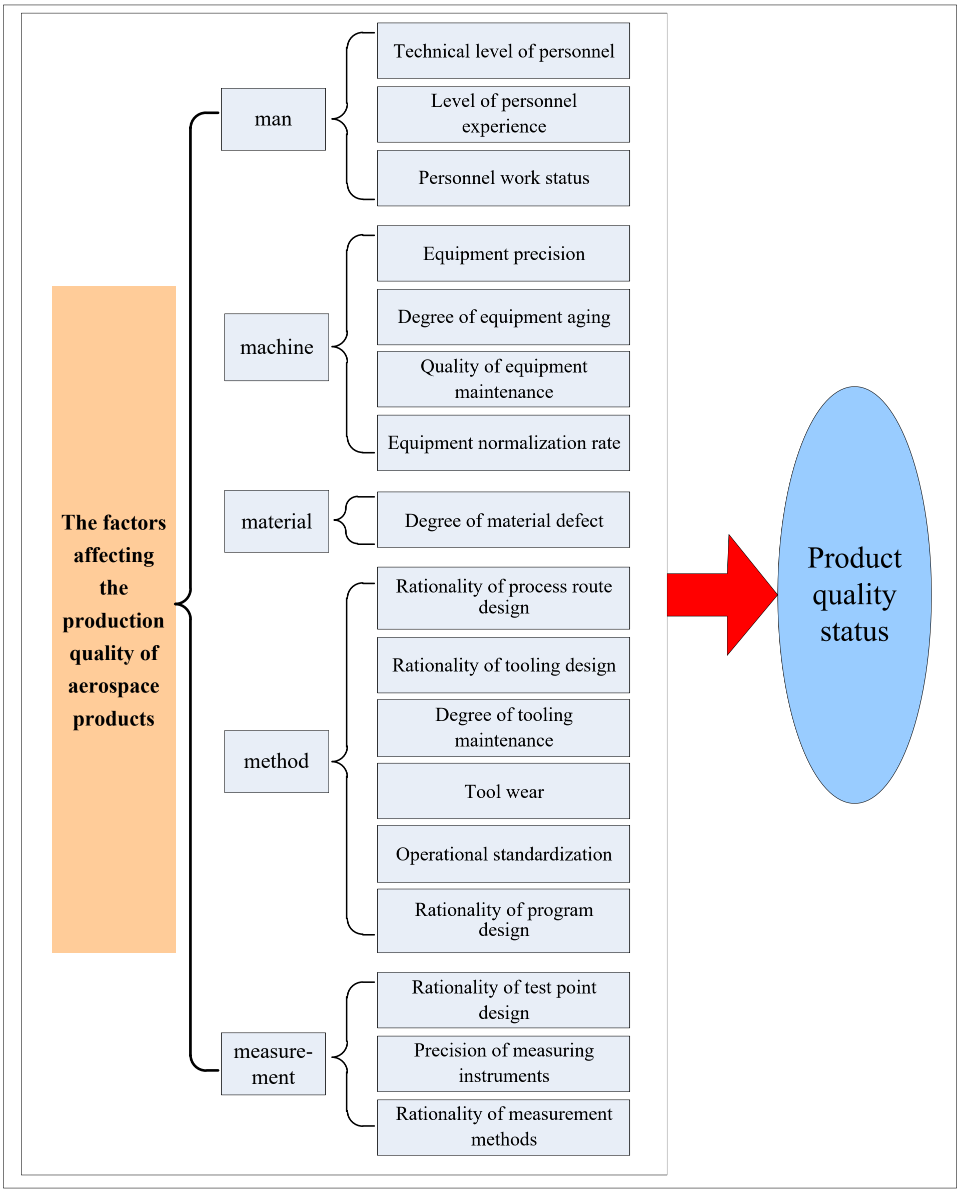


Fig. 2 The factors affecting the production quality of aerospace products

Affecting factors with regard to the machine are mainly the equipment precision, the degree of equipment aging, the quality of equipment maintenance, and the equipment normalization rate. The term "equipment precision" is used to describe the degree of refinement of equipment processing, including ordinary, computer numerical control (CNC), or precision machine tools and other equipment. There is a notable difference in precision between these categories. The degree of aging is directly proportional to the cumulative usage time of the equipment in question. Therefore, the longer the cumulative usage time of the equipment, the greater the degree of aging will be. The quality of equipment maintenance can be defined as the degree and quality to which equipment is maintained. The equipment normalization rate refers to the normal work time of the equipment/the production time of the product. The impact of the aforementioned factors on production results will vary in accordance with their differences.

The main factor affecting the material is the degree of material defects. The presence of material defects, such as cracks, impurities, and sand holes, will inevitably exert an influence on the quality of the final product to a certain extent.

Affecting factors in terms of production methods are the rationality of process route design, the rationality of tooling design, the degree of tooling maintenance, tool wear, operational standardization, and the rationality of program design. Process route refers to the processing route of the product, and rationality of tooling design refers to the operability of the process equipment for the processing of the product. The more reasonable the process route and tooling design, the greater the likelihood of successful product processing. Furthermore, the degree of tooling maintenance, tool wear, operational standardization, and the rationality of program design also have an impact on the quality of the processing.

Affecting factors with regard to the measurement are the rationality of the test point design, the precision of the measuring instruments, and the rationality of the measurement methods employed. The quality of the processing will be adversely affected by the unreasonable setting of test points, the low precision of measuring instruments, and the use of unreasonable measuring methods.

The aforementioned five affecting factors collectively determine the quality of production of aerospace products. An analysis of these factors, coupled with the identification of the most significant factors and the establishment of a correlation between them, can effectively inform the classification of production quality and facilitate decision-making.

The affecting factors and factor value settings are presented in Table 1.

Information regarding the on-site processing of a product in recent years is obtained from the MES system, as detailed in Table 2. The following table presents the actual data on some of the original affecting factors and the actual processing results.

The correlation coefficients between the original affecting factors of the production quality of aerospace products and between the original affecting factors and the production results (qualified) have been calculated and presented in Fig. 3 for purposes of illustration. The magnitude of the correlation coefficients serves to reflect the strength of the relationship between the various original affecting factors and the extent to which they impact the production outcomes. As demonstrated by the correlation coefficient calculation, 13 original affecting factors with a notable impact on production results have been identified as the original factors for subsequent dimensionality reduction. The 13 original affecting factors are as follows: the technical level of personnel (x_1), the level of personnel experience (x_2), the personnel working status (x_3), the degree of equipment aging (x_5), the equipment normalization rate (x_7), the degree of material defects (x_8), the rationality of process route design (x_9), the rationality of tooling design (x_{10}), tool wear (x_{12}), operational standardization (x_{13}), the rationality of program design (x_{14}), the rationality of test point design (x_{15}), and the rationality of measurement methods (x_{17}).

Table 1 Affecting factors and thresholds

Category	Factor	Factor name	Factor value	Description of factor
Man	x_1	Technical level of personnel	{1,2,3}	primary, intermediate and senior
	x_2	Level of personnel experience	{1,2,3}	Years of experience in this position: 1: 0-2 years 2: 3-5 years 3: more than 5 years
	x_3	Personnel working status	{1,2}	Personnel work status: 1: Bad (more than 8 hours of continuous work or more than 5 days of continuous work) 2: Good (less than 8 hours of continuous work and less than 5 days of continuous work)
Machine	x_4	Equipment precision	[0,1]	The degree of refinement of equipment processing, such as ordinary, CNC or precision machine tools, etc.
	x_5	Degree of equipment aging	[0,1]	Based on cumulative usage time of equipment
	x_6	Quality of equipment maintenance	[0,1]	Quality of equipment repair and maintenance
	x_7	Equipment normalization rate	[0,1]	Normal working time of the equipment / Production time
Material	x_8	Degree of material defects	[0,1]	Material integrity, material processing features
Method	x_9	Rationality of process route design	[0,1]	Design rationality of processing sequence and content
	x_{10}	Rationality of tooling design	[0,1]	Operability of tooling design
	x_{11}	Degree of tooling maintenance	[0,1]	maintenance level
	x_{12}	Tool wear	[0,1]	Tool wear level
	x_{13}	Operational standardization	[0,1]	Operational standardization
	x_{14}	Rationality of program design	[0,1]	Organizational rationality, operability
Measurement	x_{15}	Rationality of test point design	[0,1]	Reasonable level of test point setup
	x_{16}	Precision of measuring instruments	[0,1]	Precision of measuring instruments
	x_{17}	Rationality of measurement methods	[0,1]	Reasonableness of the test engineer's measurement method

Table 2 Each original affecting factor in MES system and actual results (partial data)

No.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	Result
1	2	3	2	0.9	0.7	0.8	0.9	0.9	0.9	0.8	0.7	0.8	0.8	0.8	0.7	0.7	0.8	1
2	3	3	2	0.8	0.6	0.7	0.6	0.8	0.6	0.7	0.8	0.9	0.8	0.8	0.8	0.5	0.9	1
3	2	2	2	0.9	0.8	0.8	0.9	0.9	0.9	0.8	0.9	0.8	0.8	0.9	0.8	0.9	0.8	1
4	1	2	2	0.7	0.7	0.7	0.9	0.8	0.9	0.9	0.6	0.7	0.9	0.9	0.9	0.7	0.8	1
5	1	3	1	0.8	0.6	0.7	0.9	0.9	0.9	0.8	0.8	0.8	0.9	0.9	0.8	0.8	0.8	1
6	3	1	2	0.9	0.6	0.9	0.9	0.9	0.9	0.8	0.6	0.9	0.8	0.9	0.9	0.9	0.9	1
7	1	1	1	0.7	0.8	0.7	0.9	0.8	0.9	0.9	0.8	0.5	0.8	0.9	0.6	0.7	0.6	-1
8	2	2	1	0.9	0.9	0.6	0.8	0.7	0.3	0.6	0.7	0.6	0.8	0.3	0.5	0.8	0.6	-1
9	2	3	1	0.8	0.8	0.8	0.9	0.7	0.7	0.2	0.8	0.4	0.8	0.9	0.6	0.9	0.5	-1
10	1	1	2	0.8	0.9	0.9	0.2	0.8	0.8	0.9	0.9	0.5	0.9	0.9	0.7	0.8	0.4	-1
11	2	3	1	0.9	0.6	0.7	0.9	0.8	0.6	0.8	0.7	0.6	0.8	0.2	0.3	0.9	0.5	-1

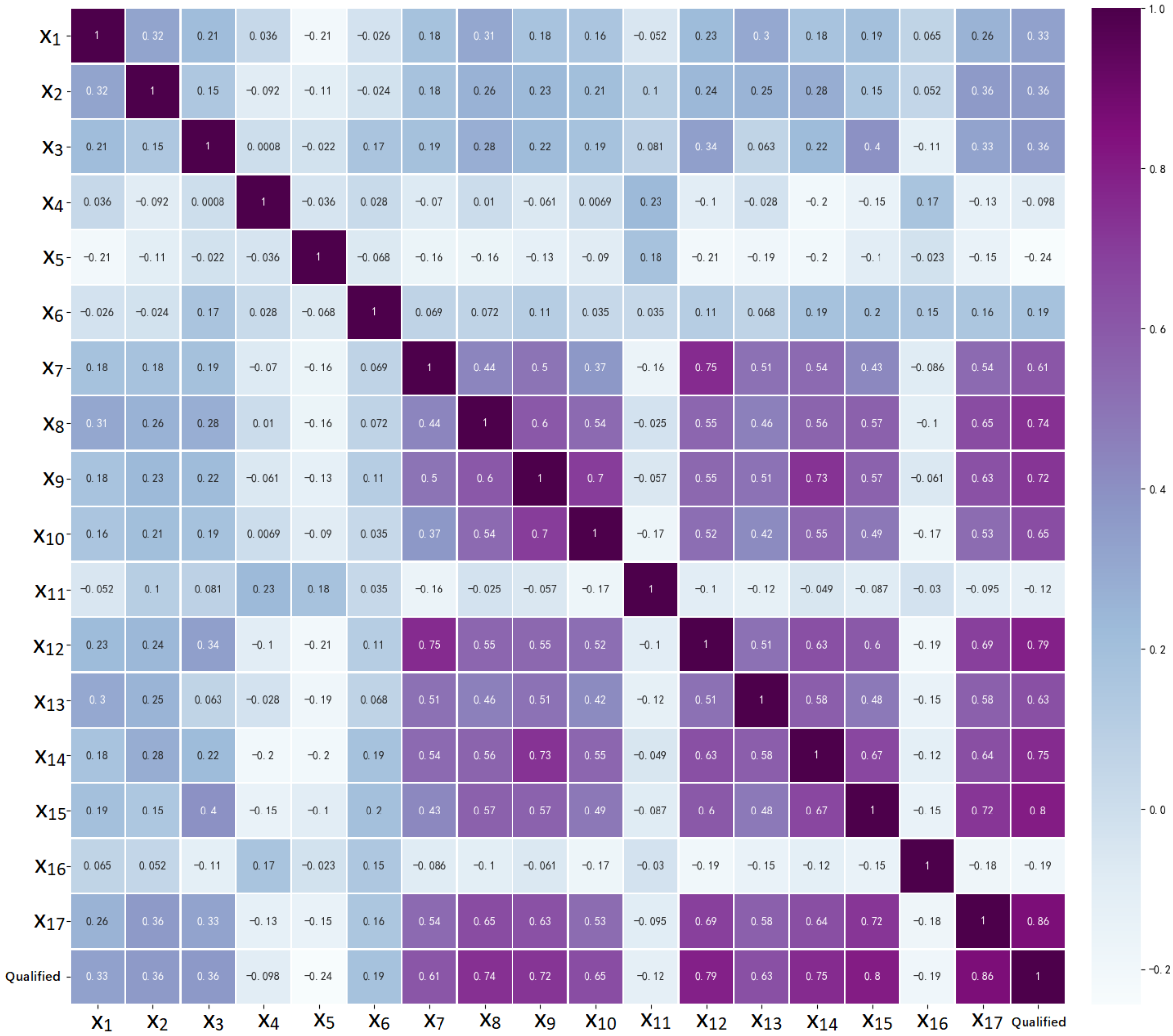


Fig. 3 The correlation coefficients between each original affecting factor and the qualified production results

3.2 Dimensionality reduction and data analysis

The geodetic distance array D_g between samples has been established, and in accordance with Eq. 2, the residual curve has been drawn and presented in Fig. 4. It is evident that as the dimensionality exceeds 6, the curve displays enhanced smoothness, with residual variations confined to a 5 % range. It is thus established that the intrinsic dimensionality of the sample is equal to 6, indicating that there are six principal features that collectively account for 95 % of the sample's features.

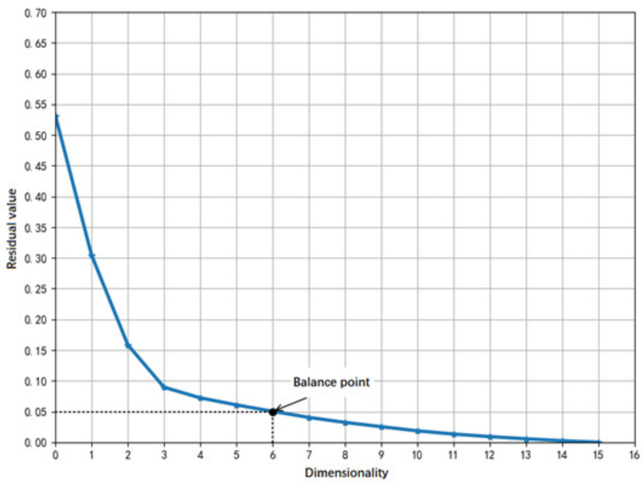


Fig. 4 Residual curve

In accordance with Eq. 4, the low-dimensional embedding result T can be calculated. This result is represented by the eigen-sample ($z_1, \dots, z_6$), which comprises six principal components after dimensionality reduction. The correlation coefficients between each principal component after dimensionality reduction and the original factors are calculated in accordance with Eq. 3 and are presented in Table 3.

The size of the correlation coefficient between each principal component and the original factors allows us to elucidate the corresponding relationship between them and to ascertain which original factors are most pertinent to the principal component after dimensionality reduction. As illustrated in Table 3, the original factors with high correlation coefficients for the first principal component z_1 include the technical level of personnel, the level of personnel experience, the personnel working status and operational standardization, which are personnel-related factors. The original factors with high correlation coefficients for the second principal component z_2 include the degree of equipment aging, equipment normalization rate and tool wear, which are equipment-related factors. The original factor with a high correlation coefficient for the third principal component z_3 is the degree of material defects, which is material-related factor. The original factors with high correlation coefficients for the fourth principal component z_4 include the rationality of process route design and the rationality of tooling design, which are process design-related factors. The original factor with a high correlation coefficient for the fifth principal component z_5 is the rationality of program design, which is program design-related factor. The original factors with high correlation coefficients for the sixth principal component z_6 include the rationality of test point design and the rationality of measurement methods, which are measurement method-related factors.

Fig. 5 illustrates the two-dimensional spatial projection of the two original factors, namely the technical level of personnel and the level of personnel experience. It can be observed that the original sample features are not readily discernible. Fig. 6 illustrates the two-dimensional spatial projection of the first feature z_1 and the second feature z_2 after dimensionality reduction using an ISOMAP-based dimensionality reduction algorithm, which can be seen that the principal components are clearly featured and can reflect the main features that affect the quality of product production.

Table 3 The correlation coefficients between each principal component after dimensionality reduction and the original factors

Principal component	x_1	x_2	x_3	x_5	x_7	x_8	x_9	x_{10}	x_{12}	x_{13}	x_{14}	x_{15}	x_{17}
z_1	0.72	0.62	0.59	-0.02	0.07	0.04	0.06	0.06	0.07	0.74	0.07	0.06	0.10
z_2	0.06	-0.07	0.07	-0.72	-0.68	0.01	-0.02	-0.02	-0.71	0.01	-0.03	0.01	-0.03
z_3	0.22	0.13	-0.09	-0.01	-0.11	-0.75	-0.11	-0.09	-0.15	-0.03	-0.12	-0.16	-0.14
z_4	0.04	0.13	0.37	0.05	-0.36	-0.13	-0.73	-0.69	-0.30	-0.28	-0.36	-0.25	-0.31
z_5	0.01	-0.01	-0.01	0.07	-0.71	0.08	0.31	0.43	-0.32	-0.04	0.86	0.19	0.06
z_6	0.01	0.01	0.04	0.01	0.25	0.01	0.25	0.20	0.06	-0.23	-0.21	-0.62	-0.62

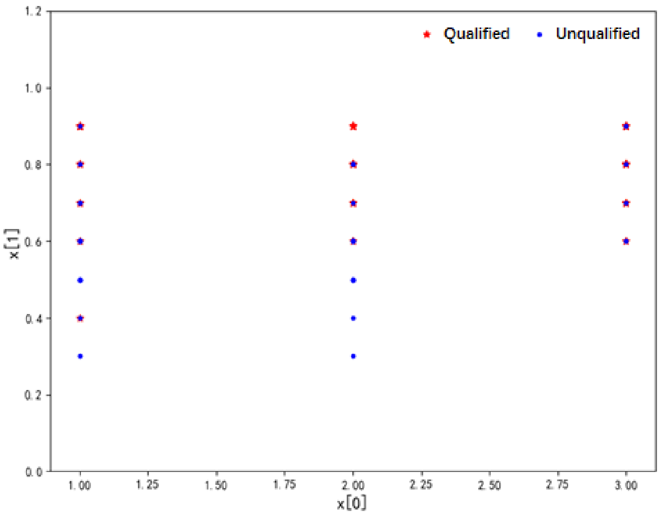


Fig. 5 2D feature space of the original factors sample

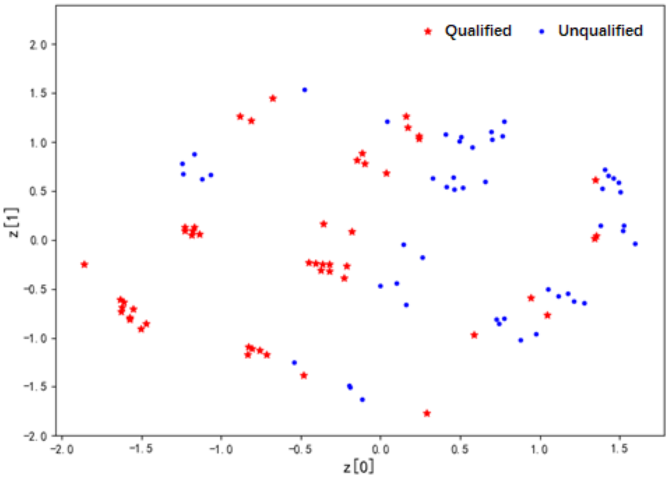


Fig. 6 2D feature space of the principal component after dimensionality reduction

3.3 Data analysis conclusion

From the correlation coefficients between each original affecting factor and the qualified production results presented in Fig. 3, as well as the correlation coefficients between the principal component after dimensionality reduction and the original factors detailed in Table 3, it can be observed that the rationality of measurement methods, the rationality of test point design, tool wear, equipment normalization rate, the degree of equipment aging, the rationality of program design, the degree of material defects, the rationality of process route design, the rationality of tooling design, the technical level of personnel, the level of personnel experience, the personnel work status, and operational standardization have a significant impact on production quality. The main features of the influence degree, in descending order of magnitude, are z_6, z_2, z_5, z_3, z_4, and z_1.

The sixth feature, z_6, pertains to the rationality of measurement methods and the rationality of test point design. To minimize the potential for error in on-site measurement results, it is recommended that standard automated measurement methods be employed wherever feasible.

The second feature, z_2, is associated with tool wear, equipment normalization rate, and the degree of equipment aging. It is recommended that regular maintenance of processing tools and equipment be conducted to ensure normalization of said tools and equipment.

The fifth feature, Z_5, pertains to the rationality of program design. It is imperative to reinforce training and review of the program design.

The third feature, Z_3, is associated with the degree of material defects and thus requires comprehensive examination through rigorous testing.

The fourth feature, Z_4, is associated with the rationality of process route design and the rationality of tooling design. It is of the utmost importance to reinforce the process route design and tooling design optimization, and these are subjected to the closest scrutiny by experts.

The first feature, z_1, pertains to the technical level of personnel, the level of personnel experience, the personnel work status, and operational standardization. It is of the utmost importance to reinforce the training and examinations of personnel.

4. Conclusion

This paper addresses the issue of identifying the main factors affecting the production quality of aerospace products. It begins by providing a comprehensive analysis of the characteristics of aerospace product production and the various types of quality-affecting factors. Secondly, the isometric feature mapping (ISOMAP) algorithm of stream learning is used to reduce the dimensionality of nonlinear data. Finally, the correlation coefficients between each principal component after dimensionality reduction and the original factors are calculated to elucidate their correspondence, and the main factors are ranked according to the degree of influence. The process can identify the factors that have the greatest impact on the quality of aerospace product production, mainly those related to inspection methods, the condition of equipment, etc. Therefore, to ensure that the quality of aerospace product production is improved, it is recommended that standard automated measurement methods be used wherever feasible. In addition, it is recommended that regular maintenance of machining tools and equipment be enhanced to ensure that they are in good condition. The proposed changes will enable the targeted use of costs to significantly improve the quality of aerospace product production and achieve cost savings. The experimental results certify the capability of the algorithm in analyzing the main factors affecting the production quality of aerospace products, eliminating redundant factor interference, highlighting the main affecting features, and obtaining the inner law of the data, with the limitation that the algorithm is only applicable to the scenarios of analysing the production quality of all kinds of aerospace products with small samples. On the basis of this study, further research and practice on online classification and prediction of aerospace product production quality will be carried out in the future.

Acknowledgement

Project supported by the Natural Science Foundation of Jiangsu Province of China (BK20222012).

References

[1] Nazifa, T.H., Ramachandran, K.K. (2019). Information sharing in supply chain management: A case study between the cooperative partners in manufacturing industry, *Journal of System and Management Sciences*, Vol. 9, No. 1, 19-47, doi: 10.33168/JSMS.2019.0102.

[2] Shafiq, S.I., Sanin, C., Szczerbicki, E., Toro, C. (2017). Towards an experience based collective computational intelligence for manufacturing, *Future Generation Computer Systems*, Vol. 66, 89-99, doi: 10.1016/j.future.2016.04.022.

[3] Liang, Q. (2020). Production logistics management of industrial enterprises based on wavelet neural network, *Journal Européen des Systèmes Automatisés*, Vol. 53, No. 4, 581-588, doi: 10.18280/jesa.530418.

[4] Păun, M.-A., Coandă, H.-G., Mincă, E., Iliescu, S.S., Duca, O.G., Stamatescu, G. (2024). Improved multi-objective genetic algorithm used to optimizing power consumption of an integrated system for flexible manufacturing, *Studies in Informatics and Control*, Vol. 33, No. 1, 27-36, doi: 10.24846/v33i1y202403.

[5] Malega, P., Daneshjo, N. (2024). Increasing the production capacity of business processes using plant simulation, *International Journal of Simulation Modelling*, Vol. 23, No. 1, 41-52, doi: 10.2507/ijsimm23-1-669.

[6] Sun, H. (2023). Optimizing manufacturing scheduling with genetic algorithm and LSTM neural networks, *International Journal of Simulation Modelling*, Vol. 22, No. 3, 508-519, doi: 10.2507/IJSIMM22-3-CO13.

[7] Yildiz, İ., Saygin, A., Çolak, S. Abut, F. (2023). Development of a neural network algorithm for estimating the makespan in jobshop production scheduling, *Tehnički Vjesnik–Technical Gazette*, Vol. 30, No. 4, 1257-1264, doi: 10.17559/TV-20220818161430.

[8] Schindlerova, V., Marcik, J., Cada, R., Sajdlerova, I. (2024). Analysis of non-conforming production in an engineering company, *Tehnički Vjesnik–Technical Gazette,* Vol. 31, No. 1, 296-302, doi: 10.17559/TV-20230907000920.

[9] Kliment, M., Trebuna, P., Pekarcikova, M., Straka, M., Trojan, J., Duda, R. (2020). Production efficiency evaluation and products' quality improvement using simulation, *International Journal of Simulation Modelling*, Vol. 19, No. 3, 470-481, doi: 10.2507/IJSIMM19-3-528.

[10] Liang, S., Chen, C., Wu, D., Chen, L., Wu, Q., Gu, T.T. (2024). An ensemble learning method for the fault multi-classification of smart meters, *Tehnički Vjesnik–Technical Gazette*, Vol. 31, No. 5, 1514-1522, doi: 10.17559/TV-20230417000543.

[11] Yin, B., Wei, X., Wang, J., Xiong, N., Gu, K. (2019). An industrial dynamic skyline based similarity joins for multi-dimensional big data applications, *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 4, 2520-2532, doi: 10.1109/TII.2019.2933534.

[12] Rosin, F., Forget, P., Lamouri, S., Pellerin, R. (2021). Impact of Industry 4.0 on decision-making in an operational context, *Advances in Production Engineering & Management*, Vol. 16, No. 4, 500-514, doi: 10.14743/apem2021.4.416.

[13] Jolliffe, I. (2002). Principal component analysis and factor analysis, In: *Principal component analysis. springer series in statistics,* Springer, New York, USA, 150-166, doi: 10.1007/0-387-22440-8_7.

[14] Cox, M., Cox, T. (2008). Multidimensional scaling, In: *Handbook of data visualization, Springer handbooks of computational statistics,* Springer*,* Berlin, Germany, 316-341, doi: 10.1007/978-3-540-33037-0_14.

[15] Wang, Z., Han, D., Li, M., Liu, H., Cui, M. (2022). The abnormal traffic detection scheme based on PCA and SSH, *Connection Science*, Vol. 34, No. 1, 1201-1220, doi: 10.1080/09540091.2022.2051434.

[16] Jiang, L., Jiang, H., Wang, H.H. (2020). Soft computing model using cluster-PCA in port model for throughput forecasting, *Soft Computing*, Vol. 24, No. 7, 14167-14177, doi: 10.1007/s00500-020-04786-y.

[17] Ye, M., Ji, C., Chen, H., Lei, L., Lu, H., Qian, Y. (2020). Residual deep PCA-based feature extraction for hyperspectral image classification, *Neural Computing and Applications*, Vol. 32, No. 7, 14287-14300, doi: 10.1007/s00521-019-04503-3.

[18] Dong, M., Yu, H., Zhang, L., Sui, Y., Zhao, R. (2023). A PCA-smo based hybrid classification model for predictions in precision agriculture, *Tehnički Vjesnik–Technical Gazette*, Vol. 30, No. 5, 1652-1660, doi: 10.17559/TV-2023053 0000682.

[19] Huber, R., Ramoser, H., Mayer, K., Penz, H., Rubik, M. (2005). Classification of coins using an eigenspace approach, *Pattern Recognition Letters*, Vol. 26, No. 1, 61-75, doi: 10.1016/j.patrec.2004.09.006.

[20] Shawe-Taylor, J., Cristianini, N. (2004). *Kernel methods for pattern analysis,* Cambridge University Press, London, UK, doi: 10.1017/CBO9780511809682.

[21] Schölkopf, B., Smola, A., Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, Vol. 10, No. 5, 1299-1319, doi: 10.10.1162/089976698300017467.

[22] Kim, K.I., Jung, K., Kim, H.J. (2002). Face recognition using kernel principal component analysis, *IEEE Signal Processing Letters*, Vol. 9, No. 2, 40-42, doi: 10.1109/97.991133.

[23] Lima, A., Zen, H., Nankaku, Y., Miyajima, C., Tokuda, K., Kitamura. T. (2004). On the use of kernel PCA for feature extraction in speech recognition, *IEICE TransactionS on Information and Systems*, Vol. E87-D, No. 12, 2802-2811.

[24] Hoffmann, H. (2007). Kernel PCA for novelty detection, *Pattern Recognition*, Vol. 40, No. 3, 863-874, doi: 10.1016/j.patcog.2006.07.009.

[25] Van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, 2579-2605.

[26] Mcinnes, L., Healy, J. Saul, N., Großberger, L. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction, *The Journal of Open Source Software*, Vol. 3, No. 29, Article No. 861, doi: 10.21105/joss.00861.

[27] Tenenbaum, J.B., de Silva, V., Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, Vol. 290, No. 5500, 2319-2323, doi: 10.1126/science.290.5500.2319.

[28] Halder, R., Fidkowski, K.J., Maki, K.J. (2024). An adaptive sampling algorithm for reduced-order models using isomap, *International Journal for Numerical Methods in Engineering*, Vol. 125, No. 8, Article No. e7427, doi: 10.1002/nme.7427.

[29] Lu, W., Shi, C., Fu, H., Xu, Y. (2023). Research on transformer fault diagnosis based on ISOMAP and IChOA-ISSVM, *IET Electric Power Applications*, Vol. 17, No. 6, 773-787, doi: 10.1049/elp2.12302.

[30] Li, M., Yang, J.H., Xu, J.W., Yang, D.B. (2009). Trend analysis method via manifold evolution in high dimensional space for state of machinery equipment, *Journal of Mechanical Engineering*, Vol. 45, No. 2, 213-218.

[31] Reddy, C.S., Reddy, M.R. (2024). Nonlinear difference subspace method of motor imagery EEG classification in brain-computer interface, *Digital Signal Processing*, Vol. 155, Article No. 104720, doi: 10.1016/j.dsp.2024.104720.

[32] Mohammed, M.A.A., Szabó, N.P., Szűcs, P. (2025). High-resolution characterization of complex groundwater systems using wireline logs analyzed with machine learning classifiers and isometric mapping techniques, *Modeling Earth Systems and Environment*, Vol. 11, Article No. 85, doi: 10.1007/s40808-024-02263-1.

APEM journal

Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 99–115
https://doi.org/10.14743/apem2025.1.530

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Privacy-preserving AI-based framework for container transportation demand forecasting in sea-rail intermodal systems

Huang, L.[a,*], **Jiang, D.Y.**[a], **Bai, T.Y.**[b]

[a]School of Economics and Management, Beijing Jiaotong University, Beijing, P.R. China
[b]China Waterborne Transport Research Institute, Beijing, P.R. China

A B S T R A C T

In response to the growing demand for accurate freight forecasting in sea-rail intermodal transportation, particularly under the constraints of stringent data protection regulations, we introduce a privacy-preserving, AI-based framework that focuses on the micro-level identification of container transport potential. The framework combines Vertical Federated Learning (VFL) with advanced feature and sample selection techniques. It leverages privacy-preserving methods, such as homomorphic encryption and random noise, enabling secure collaboration between ports and railways while safeguarding commercially sensitive data. Through extensive experiments, our framework demonstrates superior performance in predicting container transport demand, significantly improving the accuracy of resource allocation and scheduling decisions for rail operators. The framework not only ensures compliance with data protection regulations but also provides valuable insights into intermodal transportation planning, optimizing both railway operations and customer service quality. This approach offers a practical solution for improving strategic decision-making in the sea-rail intermodal sector amid increasing privacy demands and complex logistical challenges.

A R T I C L E   I N F O

Keywords:
Freight demand forecasting;
Vertical federated learning;
Privacy-preserving methods;
Sample and feature selection;
Machine learning;
Homomorphic encryption;
Resource allocation and scheduling

**Corresponding author:*
lhuang@bjtu.edu.cn
(Huang, L.)

Article history:
Received 20 October 2024
Revised 19 February 2025
Accepted 3 March 2025

1. Introduction

With the continuous expansion of global trade, sea-rail intermodal transportation has emerged as a crucial part of modern logistics networks, effectively combining the advantages of different transportation modes and optimizing the utilization of diverse transportation resources[1]. Every day, major ports and railway systems worldwide process vast amounts of transport data, covering multiple stages from ship docking and cargo handling to final rail transportation [2]. Although this data is crucial for improving transport efficiency and optimizing logistics management, its fragmented storage across various organizations (such as customs, ports, and railway companies) presents significant challenges for efficient utilization and integrated analysis [3]. This data fragmentation not only limits information sharing and flow but also increases uncertainty throughout the transport chain. For example, delays in data exchange between port and rail departments can lead to prolonged container dwell times, ultimately reducing overall transport efficiency.

Due to the fragmented nature of the data, each party has a very limited view of the complete dataset, which not only affects information flow and collaboration efficiency but also increases operational complexity. More importantly, the growing demand for data privacy protection from data holders has made it increasingly difficult to integrate and analyse these datasets together, raising concerns about potential data leaks [4]. Therefore, developing an approach that allows for the efficient integration of fragmented data without violating privacy is now a critical challenge in sea-rail intermodal data analysis.

Federated learning offers an innovative solution to these challenges by enabling multi-party data collaboration without sharing raw data [5]. The core principle of federated learning is that participants can collaboratively train models while keeping their respective data private. Vertical Federated Learning (VFL), in particular, is well-suited for scenarios where different organizations hold different data features but share the same samples [6]. The VFL framework ensures data privacy for all parties while utilizing the complementary data from multiple sources to improve model accuracy.

Securing high-quality training datasets has always been a central challenge in machine learning and AI applications. The representativeness and quality of training data directly impact model performance. However, collecting and labelling sufficient high-quality data is costly. In a federated learning system, the selection of training samples and features plays a significant role in model performance. For instance, in horizontal federated learning, low-quality data—such as incorrect labels or skewed class distributions—can result in low and unstable model accuracy [7]. In vertical federated learning, where data features are distributed across different organizations and label access is limited, the challenge of selecting effective training samples and features becomes even more complex [8].

To address these challenges, this paper introduces a VFL framework based on Gradient Upper Bound Norms and Feature Joint Information Gain. This framework aims to optimize the analysis process for identifying potential import container sources in sea-rail intermodal transportation while ensuring data privacy protection. The key concept is to assess the importance of each participant's features to the overall model by calculating information gain, considering feature interactions via joint information gain. Sample importance is measured using Gradient Upper Bound Norms, determining which samples are best suited for model training. Additionally, the training and feature selection process incorporates homomorphic encryption and random noise to ensure privacy protection during model training. This innovative framework provides new insights and practical solutions for data analysis and decision-making in the sea-rail intermodal transportation system.

The structure of this article is as follows: Section 2 presents a thorough literature review on Freight Demand Forecasting, Vertical Federated Learning (VFL), and Sample and Feature Selection methods. Section 3 discusses the critical issue of data privacy protection within the scope of identifying potential containers for sea-rail intermodal transportation and formulates the core research problem. Section 4 provides a detailed explanation of the proposed framework, including the methodology and algorithms used to address the challenges of privacy-preserving container identification. Section 5 outlines the experimental setup and presents the results, comparing the performance of our method with existing state-of-the-art algorithms. Finally, Section 6 concludes by evaluating the effectiveness of the framework and discussing the business implications for railway container identification and intermodal transportation planning.

2. Literature review

This study stands at the intersection of the research streams on Freight Demand Forecasting, Vertical Federated Learning, Sample and Feature Selection methods. We comprehensively review the previous literature in each research stream as follows.

2.1 Freight demand forecasting

In the field of freight demand forecasting, methodologies have evolved from traditional statistical approaches to more advanced models that combine multiple techniques for improved accuracy and adaptability.

Early approaches, such as time series models like ARIMA, primarily focused on leveraging historical data trends to make future predictions. Regression models followed, incorporating external variables such as economic indicators. For instance, Khan and Khan [9] applied multivariate time series methods, including the Johansen co-integration and error correction model, to capture both short- and long-run dynamics of rail freight demand. Over time, these models have been supplemented with more sophisticated machine learning techniques, such as LSTM networks, which have proven effective in handling sequential data and capturing long-term dependencies [10].

As the complexity of freight data increased, machine learning models like Random Forests and Neural Networks gained prominence. Salais-Fierro and Martínez [11] demonstrated the superior accuracy of Artificial Neural Networks (ANNs) over traditional statistical models, particularly in forecasting freight demand using historical transportation management system (TMS) data. More recently, hybrid approaches that blend machine learning with other techniques have emerged as powerful tools for freight demand forecasting. Hassan *et al*. [12] introduced a reinforcement learning framework that combines time series models and machine learning algorithms in a rolling horizon to improve prediction accuracy over various time periods.

Other hybrid models have sought to improve interpretability and predictive power by incorporating domain-specific insights. For instance, Liu *et al*. [13] combined Grey Relational Analysis (GRA) with Deep Autoencoder Neural Networks (DNN) to enhance railway freight demand prediction. Ling *et al*. [14] introduced the Spatio-Temporal Heterogeneous Graph Attention Network (STHAN), which captures both spatial and temporal relationships within freight transportation data, demonstrating the growing complexity of models designed to account for multiple data dimensions. Econometric models also remain a staple in freight demand forecasting. Lu *et al*. [15] used input-output models to examine the effects of economic growth and structural changes on freight demand, emphasizing the continued relevance of economic indicators in freight modelling.

To date, most studies have focused on macro-level freight demand forecasting, often overlooking micro-level analysis that could optimize logistics at the individual container level. This gap in the literature is particularly important for sea-rail intermodal transportation, where predicting the transport potential of individual containers is critical for optimizing resource allocation and planning. The current study addresses this gap, offering new insights into identifying freight demand for sea-rail intermodal carriers at the micro level.

2.2 Vertical federated learning

Vertical Federated Learning (VFL) is designed for scenarios where different organizations hold disjoint sets of features for the same users or entities [16]. VFL enables organizations to jointly train machine learning models while keeping their raw data private, which is essential for privacy protection.

VFL operates primarily through two architectures: Aggregation-based VFL (aggVFL) and Split-based VFL (splitVFL) [17]. In aggVFL, each party trains its local model, and the server aggregates the results to produce a global model. Tree-based models such as SecureBoost [18] and SecureGBM [19] often operate in this framework, utilizing techniques like homomorphic encryption to ensure privacy. Meanwhile, splitVFL uses a more dynamic approach where a trainable global model is split across parties, with neural network-based models being common [20]. This allows the parties to collaborate on training without exchanging sensitive label information, with only the server retaining access to the global model [21]. Neural network-based approach has proven effective across various applications, from financial systems [22, 23] to healthcare [24], ensuring data privacy while maximizing the utility of distributed datasets.

In VFL, both sample and feature selection play crucial roles in improving communication efficiency and ensuring model performance. However, traditional methods face challenges due to privacy constraints and the large communication overhead involved. In feature selection, approaches like SFFS [25] struggle with contextual dependencies and heavy parameter transmission. To address this, methods such as FedSDG-FS [26], LESS-VFL [27] focus on reducing the impact of noisy features through advanced filtering mechanisms, maintaining privacy while ensuring feature importance, though it lacks consideration of feature correlations. For sample selection, VF-PS [28] focuses on selecting a subset of important participants. The LEARN framework [29] proposes a solution by selecting representative samples without requiring full-sample training.

2.3 Sample selection and feature selection

In machine learning, sample selection and feature selection are essential for improving model performance, reducing computational costs, and avoiding overfitting. In centralized learning, feature selection methods are usually divided into three categories: filter, wrapper, and embedded methods [30]. Filter methods calculate statistical relationships between features and the target variable, e.g., Gini impurity [31], mutual information. Wrapper methods evaluate different feature subsets by iteratively training models, though this can be computationally expensive [32]. Embedded methods, such as Lasso regression, integrate feature selection within the training process [33], offering a more balanced approach between accuracy and efficiency. Mlinarič *et al.* [34] compared various classifiers (Decision Tree, Random Forest, Bagging, and Gradient Boosting) for feature selection in automated end-of-line quality inspection of electric motors.

Sample selection focuses on selecting the most representative or important data samples for training, especially useful in scenarios with large datasets or limited labels [35]. It is particularly critical in situations where computational resources are constrained or data labelling is expensive. Traditional sample selection methods include uncertainty-based selection, where the model selects samples the most uncertain about for further labelling [36], and representativeness-based selection, where clusters or core sets are used to select samples that represent the overall data distribution [37].

However, in VFL scenarios, the lack of global data visibility adds significant complexity to both sample and feature selection. Each party holds a portion of the data (either features or samples) and cannot directly share raw data due to privacy constraints, rendering centralized selection approaches impractical. While emerging methods like LESS-VFL [27] and LEARN [29] address these challenges by introducing secure and efficient communication protocols that enable local computations and selective data sharing, there is still considerable room for further research.

3. Problem formulation

3.1 Data sharing status and problems

Current data sharing in container sea-rail intermodal transportation heavily relies on the point-to-point exchange model, particularly through Electronic Data Interchange (EDI) between ports and railway stations. While this model provides a standardized and streamlined approach, it imposes strict requirements for data transmission based on specific message standards for different data types. As a result, the format of data exchanges is highly regulated, and participants must closely adhere to these standards when transmitting key data fields through interface protocols.

The data exchange process in sea-rail intermodal transport involves multiple key stakeholders, such as ports, customs, freight forwarders, shipping companies, railway operators, and final cargo recipients. As containers transition between modes of transport, such as from sea to rail, real-time information sharing becomes increasingly crucial. However, the current lack of a unified data-sharing infrastructure, combined with delays in exchanging critical information, can lead to several challenges. These include prolonged container dwell times, miscommunication, and operational inefficiencies, all of which may result in shipment delays, information asymmetry, or even cargo loss.

In summary, current data sharing in sea-rail intermodal transportation provides essential support for the basic operations of various stakeholders and plays a crucial role in ensuring coordination between operations and organizations. However, due to the need to protect commercial secrets, comply with data privacy regulations, and ensure data security, the scope and effectiveness of existing data-sharing mechanisms are limited. Participants in the sea-rail intermodal chain are unable to share all their data unconditionally.

This selective data sharing, while safeguarding commercial interests, customer privacy, and data security, greatly restricts the potential for data mining. It limits the ability to fully leverage data for improving the overall efficiency of the transportation system, forecasting logistics demand, and optimizing resource allocation. Furthermore, the current reliance on message exchanges and data interfaces poses additional security risks, such as potential interception or tampering during transmission. The delays in information transfer prevent real-time updates on the transport process, hindering timely decision-making.

Additionally, challenges such as non-uniform data formats, inconsistent data quality, and compatibility issues between different information systems further complicate the data-sharing process. These problems often require extensive data cleaning and validation to ensure accuracy, increasing both the cost and complexity of data sharing.

3.2 Potential container identification scenario

In the sea-rail intermodal import process, the railway transportation workflow includes several key stages: freight forwarder application, daily train requests, railway acceptance, scheduling approval, plan preparation, and departure confirmation. Before these steps, the railway freight marketing department typically conducts freight demand mining, which is crucial for efficient resource allocation, maximizing transport efficiency, and minimizing costs.

Currently, railway freight departments conduct market research-based freight demand mining. This process involves identifying transport demand across various regions and industries. The departments engage directly with shippers, offering freight rate subsidies to encourage them to choose rail transport for container shipments from ports. However, this method is time-consuming and labour-intensive, resulting in slow progress in increasing the sea-rail intermodal ratio and facing bottlenecks. Additionally, the current approach does not utilize big data and related technologies for data analysis, limiting the accuracy and effectiveness of freight demand mining.

As illustrated in Fig. 1, the traditional railway transport process is optimized by analysing data such as port schedules, container storage, documentation, operational records, and customs information. Based on the results of potential container identification, preliminary railway transport plans—such as block trains and direct services—are formulated. The railway freight marketing department then uses these plans and transport products to conduct targeted marketing to customers. Once the freight sources are secured, the pre-compiled plans are seamlessly integrated into the existing workflow for final plan preparation. However, identifying potential container demand in real-time and ensuring privacy requires more advanced data-sharing solutions, which are discussed in the following section.

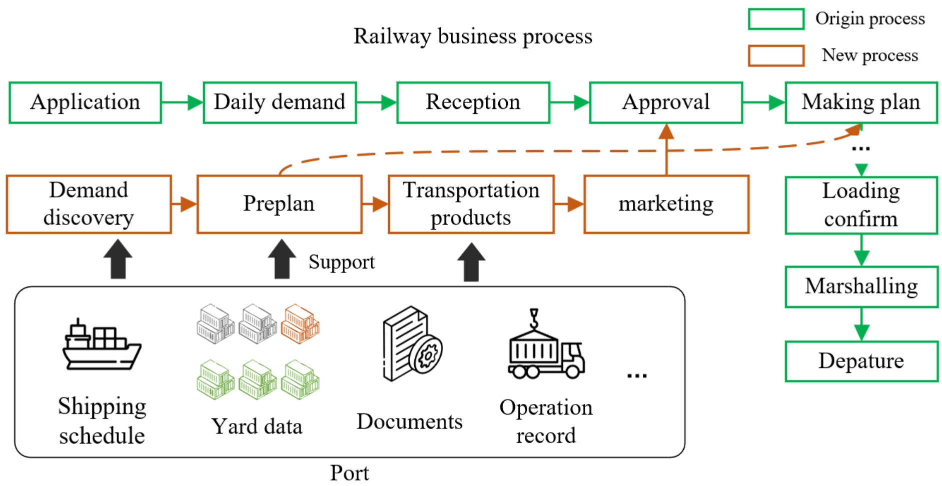


Fig. 1 Optimized railway transportation process

3.3 Privacy computing needs

The core of current data sharing in sea-rail intermodal transportation lies in supporting operational collaboration and process coordination between railway and ports, rather than indiscriminately sharing all data between both parties. The sharing mechanism focuses on improving joint operational efficiency, ensuring that data exchange enhances cargo transport efficiency, optimizes scheduling, and improves customer service. It primarily targets the exchange of essential operational data, such as train requests from ports, railway confirmations of train availability and estimated arrival times, loading confirmations, and intermediate stops.

However, the current data-sharing system is not suitable for deeper freight demand mining in sea-rail intermodal transport. Railway needs to dynamically analyse a broader range of data in real time, including ship schedules, operations, yard conditions, and destination flows of all containers. These types of raw data, however, are considered sensitive by organizations such as ports and customs and are large in volume. Under the current data-sharing framework, effective real-time sharing of this information is not possible.

The primary goal of freight demand mining in sea-rail intermodal transport is to enable railway freight marketing departments to accurately identify potential container freight demand while ensuring the security of data across multiple parties. Based on this demand, railway can dynamically optimize transport organization and train schedules. The existing data-sharing mechanism is inadequate for this purpose, requiring the development of a new solution. Techniques such as federated learning and privacy computing are necessary to allow efficient and secure information sharing, enabling the mining of potential freight demand while safeguarding sensitive data across all stakeholders in sea-rail intermodal transportation.

4. The GUBN-FJIG framework

Since the participants in sea-rail intermodal transportation, namely ports and railway, hold different features of the training samples, a vertical federated learning model is required. Existing feature selection methods typically need direct access to training data, the model training process, and labels, but this is not allowed in vertical federated learning due to privacy protection requirements. Additionally, the features held by the clients in vertical federated learning may interact with each other, and current methods tend to overlook these interactions and their joint impact on the target variable.

We proposed a vertical federated learning sample and feature selection framework based on Gradient Upper Bound Norm and Feature Joint Information Gain (GUBN-FJIG framework). It consists of three submodules: feature importance initialization, sample importance calculation, and important sample and feature selection. Fig. 2 illustrates the flowchart of this sample and feature selection framework.

In the vertical federated learning framework, the dataset of N samples is divided into M parts, denoted as $D = \{D_1, \cdots, D_M\}$, where each client holds a unique feature set $\{f_{m,1}, \cdots, f_{m,d_m}\}$ and the local sample $x_{n,m} \in \mathbb{R}^{d_m}, n \in [N]$. Typically, the server S holds the sample labels $y_n \in \mathbb{R}, n \in [c]$. With the server's coordination, all clients $m \in [M]$collaboratively contribute to the global model by sharing encrypted data to protect privacy. It allows clients to collaboratively train a global model by selecting important features and samples, while minimizing the global risk $R(\theta_s)$.

$$R(\theta_s) = \mathbb{E}_{x,y} L\left(h\left(\theta_{z_1}, z_1, \dots, z_m, y_n\right)\right) \tag{1}$$

Each client m trains a local parameter h_m, representing the local dataset $x_n^m \in \mathbb{R}^{d^m}$, which is mapped to a lower-dimensional space $z_n^m := h_m(\theta_m, x_n^m \odot s_m) \in \mathbb{R}^{d_f^m}$, where $d_f^m \ll d^m$, and $s_m = \{0,1\}^{d_m}$ indicates the selected features. The server coordinates the process by optimizing a joint model, with parameters $\theta_0 := \{w_1, \cdots, w_M, \alpha_0\}$, where $w_m \in \mathbb{R}^{d'_m}$ are the parameters of the interaction layer. These parameters are combined with the lower-dimensional embeddings z_n^m sent by each client. α_0 represents the parameters other than interaction layer.

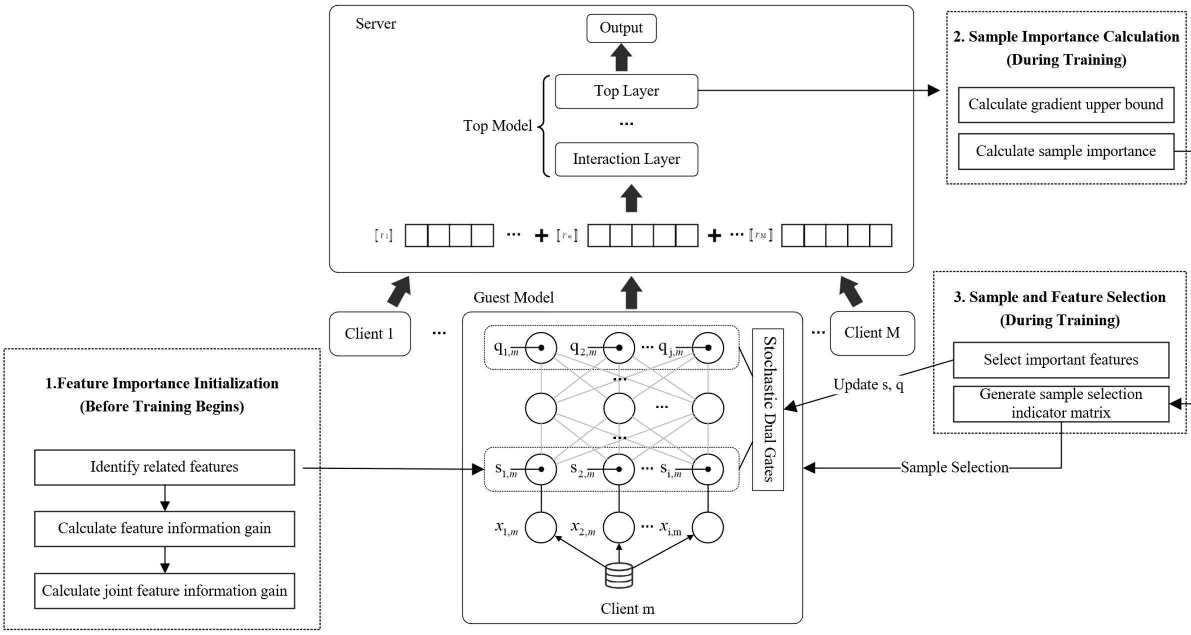


Fig. 2 GUBN-FJIG vertical federated learning framework for sample selection and feature selection

Following the Gaussian stochastic dual-gate used in FedSDG-FS [26], we utilize the l_0 norm to constrain the number of non-zero parameters in the model, minimizing the risk $R(\theta, s, q)$ to construct the global model. Due to the large variance in the Bernoulli variables s_m and q_m during feature selection optimization, a continuous relaxation based on the Gaussian distribution is applied, approximating each Bernoulli variable in s_m and q_m with parameters $\mu_{m,i}$ and $\omega_{m,j}$.

$$R(\theta, s, q) = \mathbb{E}_{x,y} L\big(h(\theta_0, r_{n,1}, \dots, r_{n,M}); y_n\big) + \lambda \sum_m (|s_m| l_0 + |q_m| l_0) \tag{2}$$

4.1 Feature importance initialization based on information gain

For the client m, if feature $f_{m,i}$ is categorical, with possible values $\left\{f_{m,j}^{(1)}, f_{m,j}^{(2)}, \dots, f_{m,j}^{(k)}\right\}$. The conditional entropy is defined as:

$$H(Y \mid f_{m,i}) = \sum_{f_{m,i}^{(k)}} p\left(f_{m,i}^{(k)}\right) H\left(Y \mid f_{m,i} = f_{m,i}^{(k)}\right) \tag{3}$$

where $p\left(f_{m,i}^{(k)}\right)$ is the probability of feature $f_{m,i}$ taking the value $f_{m,i}^{(k)}$, and $H\left(Y \mid f_{m,i} = f_{m,i}^{(k)}\right)$ is the entropy of Y given that $f_{m,i}$ takes the value $f_{m,i}^{(k)}$. The calculation of conditional entropy is as follows:

$$H\left(Y \mid f_{m,i} = f_{m,i}^{(k)}\right) = -\sum_{c \in \mathcal{C}} p\left(c \mid f_{m,i}^{(k)}\right) \log_2 p\left(c \mid f_{m,i}^{(k)}\right) \tag{4}$$

For continuous features, methods like binning or box plots can be used to discretize the feature. The information gain of a feature is calculated as $IG(Y, f_{k,i}) = H(Y) - H(Y \mid f_{k,i})$. For two features $f_{m,i}$ and $f_{m,j}$ of client m, their joint conditional entropy is calculated as $H\left(Y \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right) = -\sum_{c \in \mathcal{C}} p(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}) \log_2 p(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')})$. The information gain of joint features is $IG(Y, f_{m,i}, f_{m,j}) = H(Y) - H(Y \mid f_{m,i}, f_{m,j})$. The feature interaction information gain is calculated as $IIG(Y; f_{m,i}, f_{m,j}) = IG(Y, f_{m,i}, f_{m,j}) - IG(Y, f_{m,i}) - IG(Y, f_{m,j})$.

The above calculations are completed through client-server collaboration, as shown in Algorithm 1. In this work, we use Paillier as homomorphic encryption method to encrypt the data requiring privacy protection during computation. This method supports addition of encrypted values and multiplication of ciphertexts by constants.

Algorithm 1: Information gain-based feature importance initialization algorithm

Input: Client m， Server S
Output: Feature importance (Initialized)
Server S
Compute the class entropy $H(Y)$
Create an indicator matrix A and encrypt: $[\![A]\!] \leftarrow \mathrm{Enc}(A)$
Send the encrypted indicator matrix $[\![A]\!]$ to all clients
Client m
Based on feature $f_{m,i}$ discretize sample U into $U_1, \cdots, U_k$
For feature $f_{m,i}$ calculate $p\left(f_{m,i}^{(k)}\right) \leftarrow \frac{|U_k|}{|U|}$
Compute $\left[\!\left[p\left(c \mid f_{m,i}^{(k)}\right)\right]\!\right] \leftarrow \frac{\sum_{n \in I(U_k)} [\![A]\!]_{n,c}}{|U_k|}$
If there is a joint feature $f_{m,j}$， discretize sample U into $U_1, \cdots, U_{k'}$:
Compute $p\left(f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)$， $\left[\!\left[p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)\right]\!\right]$，
Add the encryption factor: $\left[\!\left[\epsilon_m p\left(c \mid f_{m,i}^{(k)}\right)\right]\!\right] \leftarrow \left[\!\left[p\left(c \mid f_{m,i}^{(k)}\right)\right]\!\right] \cdot \epsilon_m$，
$\left[\!\left[\epsilon_m p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)\right]\!\right] \leftarrow \left[\!\left[p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)\right]\!\right] \cdot \epsilon_m$
Send $\left[\!\left[\epsilon_m p\left(c \mid f_{m,i}^{(k)}\right)\right]\!\right]$， $\left[\!\left[\epsilon_m p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)\right]\!\right]$ to the server
Server S
Compute $\epsilon_m p\left(c \mid f_{m,i}^{(k)}\right) \leftarrow Dec\left(\left[\!\left[\epsilon_m p\left(c \mid f_{m,i}^{(k)}\right)\right]\!\right]\right)$
Decrypt $\epsilon_m p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right) \leftarrow Dec\left(\left[\!\left[\epsilon_m p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)\right]\!\right]\right)$
Compute $\log_2\left(\epsilon_m p\left(c \mid f_{m,i}^{(k)}\right)\right)$, $\log_2\left(\epsilon_m p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)\right)$ and send to the client
Client m
Remove noise $\log_2 p\left(c \mid f_{m,i}^{(k)}\right) \leftarrow \log_2\left(\epsilon_m p\left(c \mid f_{m,i}^{(k)}\right)\right) - \log_2 \epsilon_m$
$\mathrm{Log}_2\, p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right) \leftarrow \log_2\left(\epsilon_m p\left(c \mid f_{m,i}^{(k)}, f_{m,j}^{(k')}\right)\right) - \log_2 \epsilon_m$
Calculate $H\left(Y \mid f_{m,i}, f_{m,j}\right)$， $H\left(Y \mid f_{m,i}\right)$ and send to the Server
Server S
Compute information gain $IG\left(Y, f_{m,i}\right)$， $IG\left(Y, f_{m,i}, f_{m,j}\right)$ and send to the client
Client m
Compute joint information gain $IIG\left(Y, f_{m,i}, f_{m,j}\right)$, and initialize $\mu_{m,j}$
$\mu_{m,j} \propto IG\left(Y; f_{m,j}\right) + \sum_{i \neq j} IIG\left(Y; f_{m,i}, f_{m,j}\right)$

4.2 Sample importance estimation

The Gradient Upper Bound Norm is used as a sample importance indicator. Its calculation is shown below, mainly involving the input and output of the model's top layers. The computation of this norm requires only one forward pass, reducing computational costs. It provides a reasonably accurate estimate of sample importance, whereas conventional norms require both forward and backward passes through the network, making them more expensive to compute.

$$\lambda(x_n, t) = \sqrt{\left|\sum_t \beta_n^t \nabla_{\alpha_n^t} L\left(h\left(\theta_0, z_{n,1}, \cdots, z_{n,M}\right); y_n\right)\right|^2} \tag{5}$$

Here, $\lambda(x_n, t)$ represents the importance of sample x_n at iteration t, and β_n^t,α_n^t are the input and output of the top layer for sample x_n at the t-th iteration. For samples with higher gradient norms in the global model's output, the sample is assigned greater importance. Conversely, to avoid selecting samples that display unusually high values, a predefined threshold parameter $\lambda(x_n, t) \leq \delta_t$ is set, where δ_t is a user-defined parameter (e.g., the median of the sample norm distribution).

The calculation of sample importance is completed through the forward propagation process, as described in Algorithm 2. By adding random noise, the server ensures privacy protection for client data. First, the client selects a batch of samples and calculates the importance score $s_{m,i}$

based on the feature importance initialization. Encrypt intermediate results $r_{n,m}$ using Paillier before send to the Server. Then the Server adds random noise to the model's parameters ϵ_a and sends the encrypted result to the client. The client decrypts the data, removes the added noise, and sends the adjusted result $g_{n,m} + \epsilon_s$ back to the server. The server then removes the final noise ϵ_s, calculates the gradient for the top layer and completes the sample importance calculation $\lambda(x_n, t)$. If $\lambda(x_n, t) \geq \delta_t$, the sample is included for training, and a sample selection indicator matrix P is generated.

Algorithm 2：Privacy-Preserving Forward Propagation Process

Input: Client m, Server S
Output: Loss L_n, Sample importance $\lambda(x_n, t)$
Client m
Select a batch of samples $x_{n,m}$ based on the set batch size
Sample $\rho_{m,i}, \gamma_{m,j}$ from $\mathcal{N}(0, \sigma^2), i \in [d_m], j \in [d']$
Calculate $s_{m,i} = max\left(0, min(1, \mu_{m,i} + \rho_{m,i})\right)$,
$q_{m,j} = max\left(0, min(1, \omega_{m,j} + \gamma_{m,j})\right)$
Record $R_m = \sum_{i \in [d_m]} \Phi\left(\frac{\mu_{m,i}}{\sigma}\right) + \sum_{j \in [d'_m]} \Phi\left(\frac{\omega_{m,j}}{\sigma}\right)$
$z_{n,m} \leftarrow h_m(\theta_m; x_{n,m} \odot s_m), r_{n,m} = z_{n,m} \odot q_m$
Encrypt $[\![r_{n,m}]\!] \leftarrow \mathrm{Enc}(r_{n,m})$
Send $[\![r_{n,m}]\!]$ to the Server S
Server S
Add random noise to the interact layer parameters: $w'_m \leftarrow w_m + \epsilon_a$
$[\![g'_{n,m}]\!] \leftarrow [\![r_{n,m}]\!] \cdot w'_m$, add random noise ϵ_s
Send $[\![g'_{n,m} + \epsilon_s]\!]$ to client m
Client m
$g'_{n,m} + \epsilon_s \leftarrow Dec([\![g'_{n,m} + \epsilon_s]\!])$
Remove the random noise ϵ_a， $g_{n,m} + \epsilon_s \leftarrow g'_{n,m} + \epsilon_s - \epsilon_a r_{n,m}$
Send $g_{n,m} + \epsilon_s$ back to the Server S
Server S
Remove the noise $g_{n,m} = g_{n,m} + \epsilon_s - \epsilon_s$
Compute $L_n \leftarrow L(h(\alpha_0, g_{n,1}, \cdots, g_{n,M}); \mathrm{y_n})$
Obtain top layer input β_n^t, calculate $\nabla_{\alpha_n^t} L(h(\alpha_0, g_{n,1}, \cdots, g_{n,M}); \mathrm{y_n})$
Calculate $\lambda(x_n, t)$

4.3 Backpropagation update

Based on the sample selection indicator matrix, the selected data participates in training and undergoes forward propagation, followed by model updates through backpropagation. As shown in Algorithm 3, to prevent data leakage, the server adds noise ϵ_s to the gradient $\left[\!\left[\frac{\partial L_n}{\partial w_m}\right]\!\right]$ during transmission. The client decrypts the result and adjusts the gradient by a scaling factor η_s before sending it back to the server. The cumulative noise ϵ_m is recorded during this process. Server updates interaction layer parameters $w_m^{'} = w_m + \epsilon_m$ with noisy gradients. The update of client-side model requires no noise, as the server uses encrypted cumulative noise for gradient calculations $\frac{\partial L_n}{\partial g_{n,m}} \cdot w_m^{'} - [\![\epsilon_a]\!] \cdot \frac{\partial L_n}{\partial g_{n,m}}$. The server sends the updated gradient back to the client, where the client uses backpropagation to update parameters such as $\mu_m, \omega_m, \theta_m$, thereby completing feature selection with s_m, q_m and updating the client model.

Algorithm 3: Privacy-Preserving Backpropagation Process

Input: Sample loss L_n, Server learning rate η_s, Client learning rate η_m

Output: Global model

Server S

1 Compute the gradient $\left[\!\left[\frac{\partial L_n}{\partial w_m}\right]\!\right] \leftarrow \frac{\partial L_n}{\partial g_{n,m}} \cdot [\![r_{n,m}]\!], \left(\frac{\partial L_n}{\partial r_{n,m}}\right)' \leftarrow \frac{\partial L_n}{\partial g_{n,m}} \cdot w_m', \frac{\partial L_n}{\partial \alpha_0}$

2 Add random noise ϵ_s, and send $\left[\!\left[\frac{\partial L_n}{\partial w_m} + \epsilon_s\right]\!\right]$ to client m

Client m

3 $\frac{\partial L_n}{\partial w_m} + \epsilon_s \leftarrow Dec\left(\left[\!\left[\frac{\partial L_n}{\partial w_m} + \epsilon_s\right]\!\right]\right)$

4 Add random noise ϵ_m, $\left(\frac{\partial L_n}{\partial w_m} + \epsilon_s\right)' \leftarrow \frac{\partial L_n}{\partial w_m} + \epsilon_s - \frac{\epsilon_m}{\eta_s}$

5 Encrypt the noise $[\![\epsilon_a]\!] \leftarrow Enc(\epsilon_a)$ and accumulate noise $\epsilon_a \leftarrow \epsilon_a + \epsilon_m$

6 Send $\left(\frac{\partial L_n}{\partial w_m} + \epsilon_s\right)'$, $[\![\epsilon_a]\!]$ to the server

Server S

7 Remove the noise $\left(\frac{\partial L_n}{\partial w_m}\right)' \leftarrow \left(\frac{\partial L_n}{\partial w_m} + \epsilon_s\right)' - \epsilon_s$

8 Update the interaction layer parameters

$$w_m' \leftarrow w_m' - \eta_s \left(\frac{\partial L_n}{\partial w_m}\right)', \alpha_0 \leftarrow \alpha_0 - \eta_s \nabla_{\alpha_0} L_n$$

9 Compute gradients, update other layer parameters

10 Remove the noise $\left[\!\left[\frac{\partial L_n}{\partial r_{n,m}}\right]\!\right] \leftarrow \left(\frac{\partial L_n}{\partial r_{n,m}}\right)' - [\![\epsilon_a]\!] \cdot \frac{\partial L_n}{\partial g_{n,m}}$, and send to client m

Client m

11 $\frac{\partial L_n}{\partial r_{n,m}} = Dec\left(\left[\!\left[\frac{\partial L_n}{\partial r_{n,m}}\right]\!\right]\right)$, compute gradients $\frac{\partial L_n}{\partial \mu_m}, \frac{\partial L_n}{\partial \omega_m}, \frac{\partial L_n}{\partial \theta_m}$

12 Update the client model

$$\mu_m \leftarrow \mu_m - \eta_m \left(\frac{\partial L_n}{\partial \mu_m} + \lambda \frac{\partial R_m}{\partial \mu_m}\right), \omega_m \leftarrow \omega_m - \eta_m \left(\frac{\partial L_n}{\partial \omega_m} + \lambda \frac{\partial R_m}{\partial \omega_m}\right)$$

$$\theta_m \leftarrow \theta_m - \eta_m \frac{\partial L_n}{\partial \theta_m}$$

5. Framework evaluation using practical data

In this section, we apply the proposed framework in a practical scenario to identify potential containers at a port in China. We choose the metrics of accuracy and   to evaluate the performance of the proposed framework. We also compare our results against baseline models.

5.1 Data description and preprocessing

The integration of the framework faces several challenges. First, there is the issue of data integration. Since the data formats used in port and railway management systems vary, significant effort will be required for data standardization and preprocessing. Secondly, many existing systems rely on outdated infrastructure, which may not be compatible with the proposed framework and may require upgrades. Finally, collaboration among multiple stakeholders is key to ensuring smooth integration.

The data for this study were gathered from several sources. Container-related data, including basic information, shipping schedules, stack storage, and operational records, were obtained from the port's container management system. The railway transport data was collected from the China Railway Research Institute, covering two transport stations at the port, with data on daily demands, waybills, and trajectory information. Road transportation data for container trucks was sourced from Baidu Maps, utilizing truck route planning services based on primary truck models and destinations. Data collection spanned from June 2022 to July 2023. Table 1 provides a summary of the datasets, including descriptions, record counts, and ownership.

Table 1 Data sources

Dataset	Fields	Numbers	Data Owner
Container Basic Information Dataset	Container ID, Type, Size, Weight, Goods Description, Trade Type, etc.	7322158	Port System
Shipping Schedules Dataset	Estimated & Confirmed Arrival Times, Work Start & Completion Times, Departure Time, etc.	1048575	Port System
Stack Storage Dataset	Stack Entry & Departure Times, etc.	9414876	Port System
Container Operation Dataset	Destination, Dispatch Time, Mode of Transportation, etc.	485792	Port System
Container Truck Road Transportation Dataset	Transportation Distance, Fuel cost, Toll Fee, Freight Charges, Duration of Transportation, etc.	150	Baidu Maps
Container Railway Transportation Dataset	Departure & Arrival Stations, Distance Covered, Freight Charges, Discount Policy, Transportation Duration, etc.	76840	China Railway Research Institute

The proposed framework is designed to be highly adaptable to different geographic regions and logistics networks with varying data structures. It employs a flexible preprocessing pipeline that can accommodate diverse data formats and structures, allowing it to integrate and process data from different sources, such as ports and railways. The framework is capable of handling differences in data granularity, such as variations in feature sets or missing values, by applying localized feature augmentation and alignment methods. This enables the framework to function effectively across regions with distinct logistical setups or data sources.

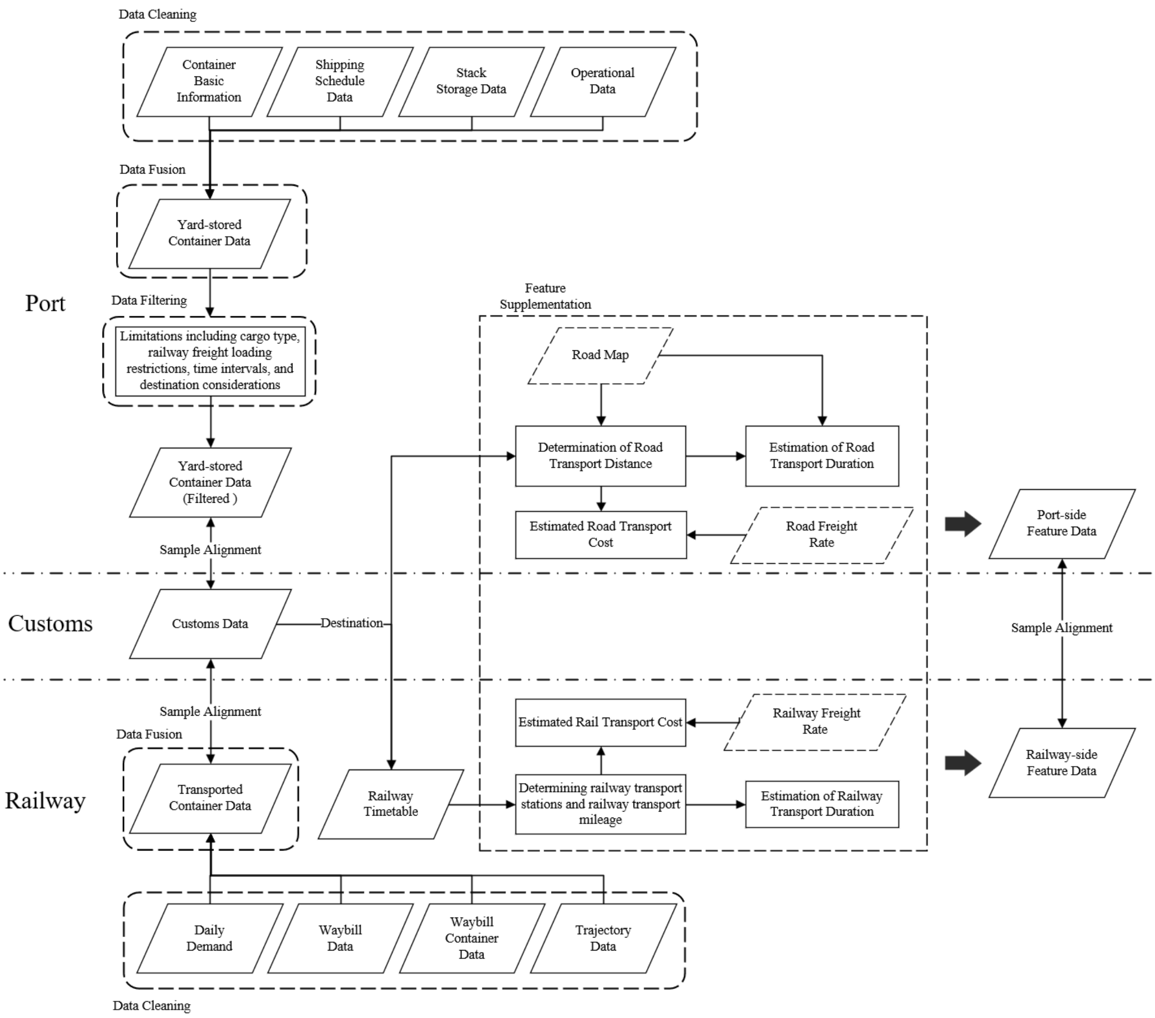


Fig. 3 Data preprocessing process in multi-party systems

All data from these three sources were integrated, as illustrated in Fig. 3. The data preprocessing in this study involves four steps: data cleaning, data integration, data filtering, and feature augmentation. The same preprocessing was applied to the port and railway datasets, though they were handled separately. After sample alignment, containers transported by road lacked railway data (distance, cost, and duration), and those transported by rail lacked road transport data. To address this, feature augmentation was applied. Missing railway features were supplemented using destination information and historical transport data, while missing road transport features were added based on destination and basic container information. Since customs hold domestic destination data, both the port and railway used pre-calculated tables for transport distance, cost, and duration for all origin-destination pairs. These were indexed using destination hashes for efficient lookup. Finally, the augmented port and railway features were aligned for consistency.

After completing the above procedures, data from multiple sources were integrated into a single dataset in logic. The dataset can be vertically divided into two parts based on the ownership of data features: port-owned features and railway-owned features, as outlined in Table 2.

Table 2 Data features and examples

Feature	Values	Type	Data Owner
Cargo weight	25.5 t, 26.3 t, …	Numeric	Port
Arrival interval	8.95 h, 6.61 h, …	Numeric	Port
Wait interval	3.13 h, 1.38h, …	Numeric	Port
Work interval	14.14 h, 6.70 h, …	Numeric	Port
Leave interval	2.80 h, 1.73 h, …	Numeric	Port
Transport interval	4.37 min, 8.20 min, …	Numeric	Port
Stack interval	249.44 h, 98.29h, …	Numeric	Port
Container type	HC, RH, FR, RF, RH, TK, …	Factor	Port
Container size	20 ft, 40 ft, …	Factor	Port
Road transportation distance	580.26 km, 149.08 km, …	Numeric	Port
Road transportation time	7.16 h, 1.93 h, …	Numeric	Port
Road fuel cost	303.58 CNY, 77.99 CNY, …	Numeric	Port
Road tolls	1047 CNY, 215 CNY, …	Numeric	Port
Road total cost	3261.04 CNY, 763.26 CNY, …	Numeric	Port
Empty container	E, F, …	Factor	Port
Trade type	D, F, …	Factor	Port
Rail transportation distance	825 km, 174 km, …	Numeric	Rail
Rail transportation time	10.31 h, 2.18 h, …	Numeric	Rail
Rail total cost	3067.6 CNY, 994.2 CNY, …	Numeric	Rail
95306 rail freight cost	3744.5 CNY, 853 CNY, …	Numeric	Rail
Discount	1439 CNY, 430.5 CNY, …	Numeric	Rail

5.2 Experimental setup

Considering that the GUBN-FJIG framework aims to identify similar transportation containers as potential freight demand, the data used for this case study was in-bureau container transport data, which has shown steady growth.

The GUBN-FJIG framework's model was trained until the prediction accuracy reached the maximum allowable iteration of 2,00. The Paillier method was used for privacy homomorphic encryption (PHE), and the Adam optimizer was applied with the learning rate and weight decay, were tuned from a grid of $\{0.01, 0.005, 0.002, 0.001\}$, and a batch size of 128, with $\lambda = 0.1$. All other hyperparameters within the network remain at their default settings.

First, the overall model's performance was compared in terms of accuracy and on the test set, to evaluate the effectiveness of feature importance initialization based on information gain within the framework, and its comparison with other feature selection methods such as all-features, Stochastic Gates (STG) and Gini impurity using similar data protection mechanisms. The same network architecture and hyperparameters were used for all methods, with the ReLU activation function and R^2.

After completing the evaluation of feature selection methods, the effectiveness of the sample selection strategy based on the gradient upper bound norm in the framework was validated by comparing the model's training efficiency and accuracy on the test set with and without a sample selection strategy.

5.3 Results and discussion

In the case study, 5-fold cross-validation was used to evaluate the impact of different feature selection methods on model performance. The dataset was divided into five subsets, with each subset used as a validation set while the remaining four were used for training. The average performance across all five folds was taken as the final evaluation metric to reduce bias introduced by data splitting and ensure the stability and reliability of the results.

Fig. 4a shows the change in training loss for different feature selection methods during training. The FJIG method achieved the fastest initial decline in training loss and eventually reached the lowest final training loss, indicating its high efficiency and good overall model performance. Its feature selection process effectively filtered out irrelevant features, allowing the model to focus on more valuable ones, improving training efficiency. Fig. 4b shows the average validation accuracy of the five-fold cross-validation using different feature selection methods (including no feature selection). STG, Gini, and FJIG methods are compared in terms of average accuracy. The results show that the FJIG method achieved significantly better validation accuracy than the other methods, especially after epoch 100, where its accuracy remained stable with less fluctuation. In contrast, the testing accuracy of the all-features method was significantly lower than other methods.

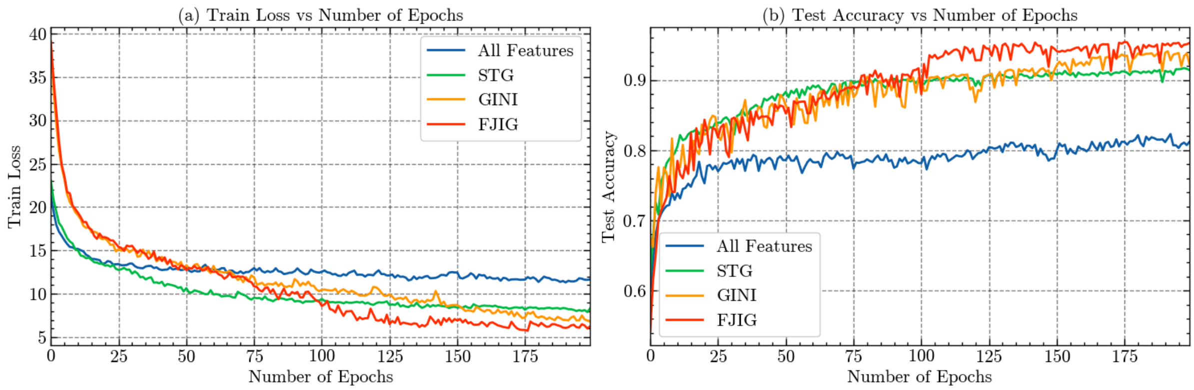


Fig. 4 Training loss and test accuracy of different feature selection methods over training epochs

Fig. 5 shows the average R^2 values across different feature selection methods during the 5-fold cross-validation. The results indicate that the FJIG method consistently maintained the highest R^2 value throughout the training process. In particular, towards the later stages of training, FJIG's R^2 value stabilized at a high level, significantly outperforming other feature selection methods. This suggests that the FJIG method, by removing noisy features, better fits the data and improves the model's predictive ability. Both STG and GINI also performed well in terms of R^2 but slightly lagged behind FJIG. The R^2 value for the all-features selection method was significantly lower than the other methods, indicating that it struggled to effectively utilize the features, especially in the presence of many noisy features.

Fig. 6 shows the changes in the number of selected features during training for different feature selection methods. The FJIG method quickly reduced the number of features early in the training process and stabilized at the minimum number of features towards the later stages. In contrast, the STG and GINI methods selected slightly more features than the FJIG method. The ability of the FJIG method to maintain high model performance with fewer features demonstrates its effectiveness in the feature selection process.

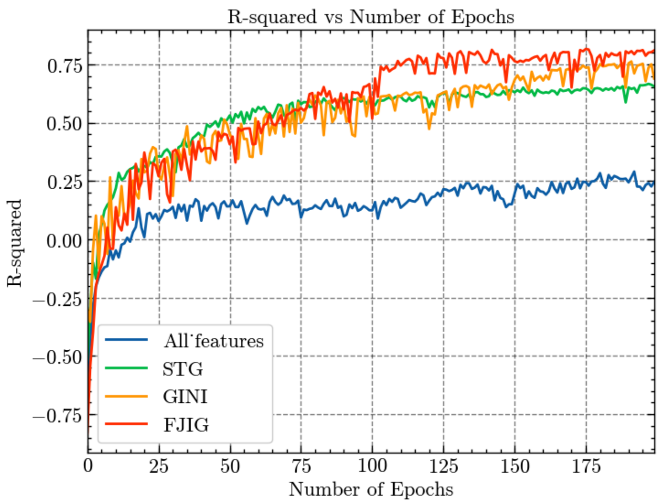


Fig. 5 R-squared of different feature selection methods over training epochs

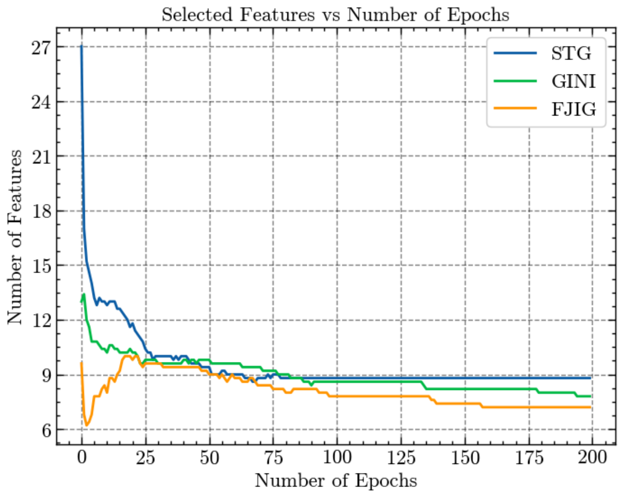


Fig. 6 Number of selected features by different feature selection methods over training epochs

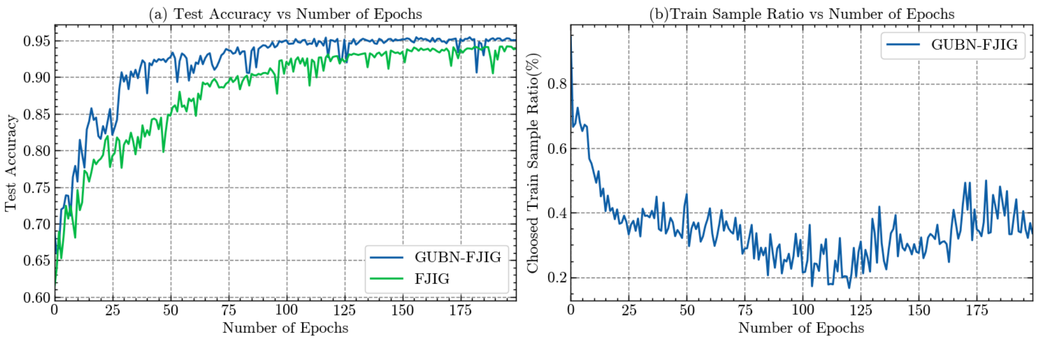


Fig. 7 Test Accuracy and Training Sample Selection Ratio of GUBN-FJIG Method over Training Epochs

Fig. 7a shows the changes in test accuracy during training for the GUBN-FJIG method with and without sample selection. In the early stages of training (around the first 50 epochs), the test accuracy of the GUBN-FJIG method increased rapidly and gradually stabilized around 0.95. In contrast, the FJIG method's test accuracy was slightly lower throughout the training process, stabilizing between 0.90 and 0.95. The superior performance of the GUBN-FJIG method in terms of test accuracy indicates that its sample selection effectively improved the model's ability to generalize to unseen test data.

Fig. 7b shows the proportion of selected training samples as the training progresses for the GUBN-FJIG method. In the early stages of training, the proportion of selected samples gradually decreased, likely because the model had not yet converged, making sample importance judgments less stable, leading to more samples being excluded. As training progressed, the proportion of selected samples stabilized and slightly increased towards the later stages, with the final selection rate stabilizing at around 40 %. This demonstrates that the GUBN-FJIG method can dynamically adjust the number of training samples involved in the process. Reducing the number of training samples in the early stages may help accelerate model convergence. In the later stages, slightly increasing the number of selected samples ensures that the model is exposed to enough information near convergence, further optimizing performance. In conclusion, the GUBN-FJIG method enhances training efficiency and generalization performance by effectively selecting training samples. The dynamic changes in the sample selection ratio reflect the method's advantage in evaluating and adapting to sample importance at different stages of training.

6. Conclusion

In this study, we proposed the GUBN-FJIG framework, which combines Gradient Upper Bound Norms (GUBN) and Feature Joint Information Gain (FJIG) for effective sample and feature selection in container transportation demand forecasting. Our approach addresses the challenges of identifying potential freight containers in the sea-rail intermodal transportation system while ensuring data privacy and computational efficiency.

Through extensive experiments, we demonstrated that the GUBN-FJIG method significantly improves model performance in terms of both accuracy and efficiency. By dynamically selecting important samples during training and filtering out irrelevant features, the method accelerates model convergence, reduces overfitting, and enhances the model's generalization ability. Our results showed that GUBN-FJIG consistently outperformed other feature selection methods, such as STG and GINI, especially in scenarios with noisy features and large datasets.

Moreover, the GUBN-FJIG method’s dynamic adjustment of the number of training samples during different stages of training contributed to its superior performance. By selecting fewer samples in the early stages to speed up convergence and increasing the sample size near convergence, the model maintained a balance between training efficiency and final performance.

In conclusion, the GUBN-FJIG framework offers a robust solution for container transportation demand forecasting in sea-rail intermodal systems. It not only optimizes model performance but also ensures data privacy protection through privacy-preserving techniques such as homomorphic encryption and random noise. From a business perspective, the framework enhances the ability of railway operators to more accurately identify potential container freight demand, leading to more informed decision-making for resource allocation and capacity planning. This results in improved operational efficiency, reduced transportation costs, and better coordination between sea and rail modes, ultimately improving service reliability and customer satisfaction. Future work could explore other scenarios of intermodal transportation systems and further improving the feature and sample selection strategies for even more efficient training and prediction.

Acknowledgements

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 52172311) and the Fundamental Research Funds for the Central Universities (Grant No. 2024JBZX042).

References

[1] Hou, D.N., Liu, S.C. (2024). Optimization of cold chain multimodal transportation routes considering carbon emissions under hybrid uncertainties, *Advances in Production Engineering & Management,* Vol. 19, No. 3, 315-332, doi: 10.14743/apem2024.3.509.

[2] Muñuzuri, J., Onieva, L., Cortés, P., Guadix, J. (2020). Using IoT data and applications to improve port-based intermodal supply chains, *Computers & Industrial Engineering*, Vol. 139, Article No. 105668, doi: 10.1016/j.cie.2019.01.042.

[3] Song, D. (2021). A literature review, container shipping supply chain: Planning problems and research opportunities, *Logistics*, Vol. 5, No. 2, Article No. 41, doi: 10.3390/logistics5020041.

[4] Lee, C., Ahmed, G. (2021). Improving IoT privacy, data protection and security concerns, *International Journal of Technology, Innovation and Management (IJTIM)*, Vol. 1, No. 1, 18-33, doi: 10.54489/ijtim.v1i1.12.

[5] McMahan, B., Moore, E., Ramage, D., Hampson, S., Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data, In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics,* Fort Lauderdale, USA, 1273-1282.

[6] Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y.-Q., Yang, Q. (2024). Vertical federated learning: Concepts, advances, and challenges, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 36, No. 7, 3615-3634, doi: 10.1109/TKDE.2024.3352628.

[7] Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., Zhang, W. (2023). A survey on federated learning: Challenges and applications, *International Journal of Machine Learning and Cybernetics*, Vol. 14, No. 2, 513-535, doi: 10.1007/s13042-022-01647-y.

[8] Khan, A., Thij, M.T., Wilbik, A. (2023). Vertical federated learning: A structured literature review, *ArXiv,* doi: 10.3934/mbe.2023122.

[9] Khan, M.Z., Khan, F.N. (2020). Estimating the demand for rail freight transport in Pakistan: A time series analysis, *Journal of Rail Transport Planning & Management*, Vol. 14, Article No. 100176, doi: 10.1016/j.jrtpm.2019.100176.

[10] Zhao, L., Cao, N., Yang, H. (2023). Forecasting regional short-term freight volume using QPSO-LSTM algorithm from the perspective of the importance of spatial information, *Mathematical Biosciences and Engineering*, Vol. 20, No. 2, 2609-2627, doi: 10.3934/mbe.2023122.

[11] Salais-Fierro, T.E., Martínez, J.A.S. (2022). Demand forecasting for freight transport applying machine learning into the logistic distribution, *Mobile Networks and Applications*, Vol. 27, No. 5, 2172-2181, doi: 10.1007/s11036-021-01854-x.

[12] Hassan, L.A.H., Mahmassani, H.S., Chen, Y. (2020). Reinforcement learning framework for freight demand forecasting to support operational planning decisions, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 137, Article No. 101926, doi: 10.1016/j.tre.2020.101926.

[13] Liu, C., Zhang, J., Luo, X., Yang, Y., Hu, C. (2023). Railway freight demand forecasting based on multiple factors: Grey relational analysis and deep autoencoder neural networks, *Sustainability*, Vol. 15, No. 12, Article No. 9652, doi: 10.3390/su15129652.

[14] Ling, S., Yu, Z., Cao, S., Zhang, H., Hu, S. (2023). STHAN: Transportation demand forecasting with compound spatio-temporal relationships, *ACM Transactions on Knowledge Discovery from Data*, Vol. 17, No. 4, 1-23, doi: 10.1145/3565578.

[15] Lu, C., Fu, S., Fang, J., Huang, J., Ye, Y. (2021). Analysis of factors affecting freight demand based on input-output model, *Mathematical Problems in Engineering*, Vol. 2021, No. 1, Article No. 5581742, doi: 10.1155/2021/5581742.

[16] Yang, Q., Liu, Y., Chen, T., Tong, Y. (2019). Federated machine learning: Concept and applications, *ACM Transactions on Intelligent Systems and Technology,* Vol. 10, No. 2, Article No. 12, doi: 10.1145/3298981.

[17] Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y.-Q., Yang, Q. (2022). Vertical federated learning: Concepts, advances, and challenges, *IEEE Transactions on Knowledge and Data Engineering,*Vol. 36, No. 7, 3615-3634, doi: 10.1109/TKDE.2024.3352628.

[18] Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D.; Yang, Q. (2021). Secureboost: A lossless federated learning framework, *IEEE Intelligent Systems*, Vol. 36, No. 6, 87-98, doi: 10.1109/MIS.2021.3082561.

[19] Feng, Z., Xiong, H., Song, C., Yang, S., Zhao, B., Wang, L., Chen, Z., Yang, S., Liu, L., Huan, J. (2019). SecureGBM: Secure multi-party gradient boosting, In: *Proceedings of 2019 IEEE International Conference on Big Data (Big Data),* Los Angeles, USA, 1312-1321, doi: 10.1109/BigData47090.2019.9006000.

[20] Romanini, D., Hall, A.J., Papadopoulos, P., Titcombe, T., Ismail, A., Cebere, T., Sandmann, R., Roehm, R., Hoeh, M.A. (2021). Pyvertical: A vertical federated learning framework for multi-headed splitNN, *ArXiv,* doi: 10.48550/arXiv.2104.00489.

[21] Fu, C., Zhang, X., Ji, S., Chen, J., Wu, J., Guo, S., Zhou, J., Liu, A.X., Wang, T. (2022). Label inference attacks against vertical federated learning, In: *Proceedings of 31st USENIX Security Symposium,* Boston, USA, 1397–1414.

[22] He, H., Wang, Z., Jain, H., Jiang, C., Yang, S. (2023). A privacy-preserving decentralized credit scoring method based on multi-party information, *Decision Support Systems*, Vol. 166, Article No. 113910, doi: 10.1016/j.dss.2022.113910.

[23] Wang, Z., Xiao, J., Wang, L., Yao, J. (2024). A novel federated learning approach with knowledge transfer for credit scoring, *Decision Support Systems*, Vol. 177, Article No. 114084, doi: 10.1016/j.dss.2023.114084.

[24] Xu, X., Peng, H., Bhuiyan, M.Z.A., Hao, Z., Liu, L., Sun, L., He, L. (2022). Privacy-preserving federated depression detection from multisource mobile health data, *IEEE Transactions on Industrial Informatics*, Vol. 18, No. 7, 4788-4797, doi: 10.1109/TII.2021.3113708.

[25] Pan, F., Meng, D., Zhang, Y., Li, H., Li, X. (2020). Secure federated feature selection for cross-feature federated learning, In: *Proceedings of NeurIPS Workshop SpicyFL,* 1-12.

[26] Li, A., Huang, J., Jia, J., Peng, H., Zhang, L., Tuan, L.A., Yu, H., Li, X.-Y. (2023). Efficient and privacy-preserving feature importance-based vertical federated learning, *IEEE Transactions on Mobile Computing,* Vol. 23, No. 6, 7238-7255, doi: 10.1109/TMC.2023.3333879.

[27] Castiglia, T., Zhou, Y., Wang, S., Kadhe, S., Baracaldo, N., Patterson, S. (2023). LESS-VFL: Communication-efficient feature selection for vertical federated learning, *ArXiv,* doi: 10.48550/arXiv.2305.02219.

[28] Jiang, J., Burkhalter, L., Fu, F., Ding, B., Du, B., Hithnawi, A., Li, B., Zhang, C. (2022). VF-PS: How to select important participants in vertical federated learning, efficiently and securely?, In: *Proceedings of 36th Conference on Neural Information Processing Systems (NeurIPS 2022),* New Orleans, Louisiana, USA, 2088-2101.

[29] Liu, T., Lyu, F., Ma, J., Deng, Y., Chen, J., Tan, Q., Zhang, Y. (2023). LEARN: Selecting samples without training verification for communication-efficient vertical federated learning, In: *Proceedings of GLOBECOM 2023-2023 IEEE Global Communications Conference,* Kuala Lumpur, Malaysia, 1217-1222, doi: 10.1109/GLOBECOM54140.2023.10437557.

[30] Khaire, U.M., Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review, *Journal of King Saud University - Computer and Information Sciences*, Vol. 34, No. 4, 1060-1073, doi: 10.1016/j.jksuci.2019.06.012.

[31] Li, X., Dowsley, R., De Cock, M. (2021). Privacy-preserving feature selection with secure multiparty computation, *ArXiv,* doi: 10.48550/arXiv.2102.03517.

[32] Roy, D., Murty, K.S.R., Mohan, C.K. (2015). Feature selection using deep neural networks, In: *Proceedings of 2015 International Joint Conference on Neural Networks (IJCNN),* Killarney, Ireland, 1-6, doi: 10.1109/IJCNN.2015.7280626.

[33] Huang, Y., Jin, W., Yu, Z., Li, B. (2020). Supervised feature selection through deep neural networks with pairwise connected structure, *Knowledge-Based Systems*, Vol. 204, Article No. 106202, doi: 10.1016/j.knosys.2020.106202.

[34] Mlinarič, J., Pregelj, B., Boškoski, P., Dolanc, G., Petrovčič, J. (2024). Optimization of reliability and speed of the end-of-line quality inspection of electric motors using machine learning, *Advances in Production Engineering & Management*, Vol. 19, No. 2, 182-196, doi: 10.14743/apem2024.2.500.

[35] Xu, X., Liang, T., Zhu, J., Zheng, D., Sun, T. (2019). Review of classical dimensionality reduction and sample selection methods for large-scale data processing, *Neurocomputing*, Vol. 328, 5-15, doi: 10.1016/j.neucom.2018.02.100.

[36] Wang, R., Kwong, S. (2010). Sample selection based on maximum entropy for support vector machines, In: *Proceedings of 2010 International Conference on Machine Learning and Cybernetics*, Qingdao, China, 1390-1395, doi: 10.1109/ICMLC.2010.5580848.

[37] Tipton, E. (2013). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments, *Evaluation Review*, Vol. 37, No. 2, 109-139, doi: 10.1177/0193841X13516324.

APEM journal

Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 116–130
https://doi.org/10.14743/apem2025.1.531

ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Enhancing racking stiffness in tall timber buildings using double-skin façades: A numerical investigation

Kozem Šilih, E.[a,*], **Premrov, M.**[a]

[a]University of Maribor, Faculty of Civil Engineering, Transportation Engineering and Architecture, Maribor, Slovenia

A B S T R A C T

The main goal of the paper is to present possible benefits in application of previously developed innovative in-plane load-bearing timber double-skin façade elements (DSF) as additional bracing elements in tall timber buildings. Therefore, a six-storey prefabricated timber structure of a height of 15 m and with a regular floor-plan is analysed by a seismic excitation of a_g = 0.225·g with a strong asymmetrical position of transparent façade elements. Two structural solutions are analysed: a hybrid system combining CLT and Light Timber-Framed walls and a non-hybrid structure made entirely of CLT. In both cases, DSF elements are first considered non-resisting and later as racking-resisting bracing elements. Numerical results show that using racking-resisting DSF elements in a hybrid system (CLT+LTF) achieves a similar increase in overall racking stiffness as a non-hybrid CLT structure with non-resisting DSF. Previous studies highlight hybrid timber systems as the preferred approach due to structural, energy-efficient, and ecological advantages. This finding is significant, offering practical benefits and new design opportunities for modern tall timber buildings with asymmetrical transparent façades, improving both energy efficiency and interior illumination in contemporary prefabricated structures.

A R T I C L E I N F O

Keywords:
Timber;
Glass;
Double-skin façades;
Racking resistance;
Mathematical modelling;
Tall timber buildings;
Seismic design

**Corresponding author:*
miroslav.premrov@um.si
(Premrov, M.)

Article history:
Received 24 July 2024
Revised 4 March 2025
Accepted 17 March 2025

1. Introduction

Due to the well-known environmental problem of harmful emissions of greenhouse gases, the profession is intensively seeking solutions in building design that will be as environmentally friendly as possible, that will produce the lowest possible CO_2 emissions, and that will also provide the highest possible standard of living comfort. Wood as a natural material has by far the best characteristics in environmental terms compared to other construction materials, as it is a CO_2-neutral material [1]. As a result, due to increased urbanisation and the concomitant need for environmentally friendly construction, there is an intense trend towards multi-storey timber buildings (MSTB), particularly in urban environments [2]. There are, of course, many limitations in this respect, particularly in terms of construction, since the modulus of elasticity of the timber elements is relatively low and therefore, particularly in areas of high wind or seismic activity, causes large horizontal displacements of the structure, which in most cases can exceed the values prescribed by the standards [3].

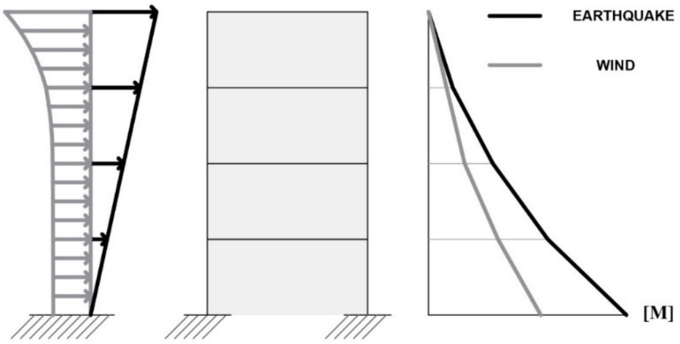


Fig. 1 Display of how wind and earthquake load increases with the height of a building at a certain location [2]

At the same time, to maximise living comfort, the profession has recently accelerated its efforts to use as much transparent glazing as possible, especially around the perimeter of the building. Such glazed surfaces allow for increased natural lighting of the interior living spaces and, on the other hand, also maximise solar heat gain, which can significantly reduce the energy demand for heating buildings during the heating season [4]. Of course, due to the increased solar radiation, such transparent elements are mostly located on the south side of the façade, i.e. they are rather asymmetrically distributed around the building perimeter. However, since such glazed elements are mostly considered in structural analysis as non-load-bearing in their plane to the action of horizontal loads, the asymmetrical plan layout of such glazed elements can results in high torsional loads at the levels of the individual storeys in the case of seismic loads, which are particularly acute in the case of multi-storey buildings, as schematically illustrated in Fig. 1 for the case of the action of two primary horizontal loads (wind and earthquake). The in-plane resistance and stiffness of such transparent elements are also not implemented in any standards yet.

Therefore, in such cases special diagonal bracing systems or other bracing solutions with common internal Light Timber-Framed (LTF) wall elements have to be incorporated into the structure of the building to satisfy all prescribed resisting requirements prescribed by the Eurocodes. However, all such structural solutions are visible and usually also not environmentally friendly and cannot contribute to any improved living comfort or they are sometimes not accepted by the architects at all. In view of the desire to provide a solution that would be at the same time optimal from in a sense of environmental performance and indoor living comfort, but also ensuring satisfactory structural resistance, transparent elements with single-panel glazing fixed to a timber frame were first developed. Such load-bearing timber-glass elements are referred to as single-skin façade (SSF) elements. However, from many experimental [5-9] and numerical studies [10, 11] it was conducted that by using only single-skin timber-glass wall elements, especially the racking stiffness did not increase in the expected manner and was not in the same range as LTF elements with the classical sheathing boards, such as OSB or fibre-plaster boards (FPB). Therefore, in this case only a relatively small additional contribution of such transparent façade elements to the overall racking stability of a whole building was achieved.

Consequently, special double-skin façade (DSF) timber-glass wall elements were further developed, first by a wide experimental study [12] and followed by a specially developed linear-elastic spring Finite Element Model [13] analysing the influence of various parameters which most significantly effect on the racking stiffness of such wall elements. The results of the numerical study were implemented first in the case of a three-storey LTF building [14]. However, in this study the position of the DSF load-bearing elements were limited to the south side of the building only. The results of the study showed a satisfactory contribution to the increased horizontal stiffness of the whole building and also to the reduction of torsional effects, especially in the first storey of the building [14].

However, there is still an important question of the applicability of such load-bearing DSF elements in much taller prefabricated timber buildings, and also with more asymmetrical position of transparent areas around the building envelope. Therefore, in our analysis, a six-storey prefabricated timber building is analysed, where DSF elements are considered as structurally non-load-bearing in the first case and as load-bearing in the second case to judge on the influence of the horizontal load impacts, with the primary purpose of analysing the influence on the reduced distortion of the building and also on the increased overall racking stiffness. Additionally, this is also the first study where the influence of resisting DSF elements is tested on a hybrid timber structural system and not only in one load-bearing system. The selection of a suitable structural system, and the energy efficiency concept strongly namely depend on the specific features of the location, particularly climate conditions, wind exposure and seismic hazard [15-17]. To satisfy in an optimal way simultaneously structural, energy and ecological aspects the choice of a hybrid timber structure seems usually to be the most favoured approach [2, 17].

Respecting this fact, therefore, in our analysis, first a hybrid structural solution (CLT+LTF) will be performed. Due to the possible distortion effects when using non-resisting transparent façade elements and which are mostly summarized on the building envelope, the CLT elements are placed on the envelope of the building, while for internal walls less stiff LTF elements are first used. In the second case the LTF internal wall elements are replaced with the CLT elements to increase the overall racking stiffness of the selected six-storey timber building. The second goal of the performed analysis is further to investigate the influence of the load-bearing DSF elements in relation to the different basic structural systems of prefabricated timber buildings. The aim of our study is to identify potentials in designing tall, prefabricated timber buildings using different structural systems with a strong asymmetrical position of the transparent façade elements around the building envelope. The influence of additional racking resistance of any transparent timber-glass wall elements is currently also not covered with any standards [18]. The obtained results would significantly improve the energy performance of modern timber buildings, as well as the indices of living comfort due to an increased illuminance. Thus, they could open many new perspectives in designing contemporary tall timber buildings which is currently somehow limited because of timber mechanical properties [2, 17].

The content of the paper is systematically organized starting with all necessary presented theoretical backgrounds in Section 2, mathematical modelling of all prefabricated LTF elements used in the study in Section 3, numerical case study on a specially selected prefabricated timber building in Section 4 and with the most important conclusions presented in Section 5.

2. Theoretical backgrounds

2.1 Main structural systems in timber buildings

Timber structural systems differ from each other in the appearance of the structure, and in the approach to planning and designing a particular system. As presented in [15] and [4], structural systems of timber buildings can be classified into six main systems:

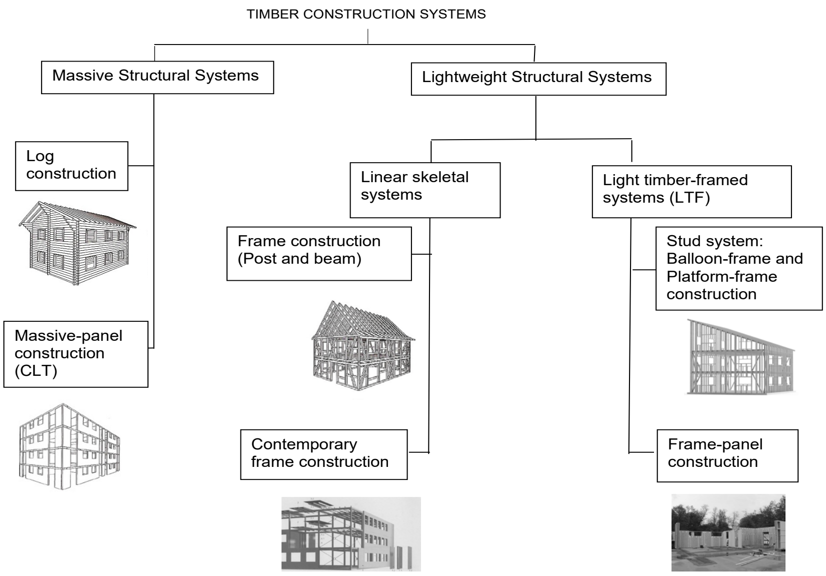


Fig. 2 Main timber structural systems [4]

- Log construction,
- Solid timber construction (CLT),
- Timber-frame construction,
- Contemporary frame construction,
- Light timber-framed construction (LTF); Balloon- and platform-frame construction,
- Light timber-framed construction (LTF); Frame-panel construction.

All systems are schematically presented in Fig. 2. However, it should be pointed out, that only the LTF Frame-panel construction and Solid timber construction as CLT are prefabricated and therefore will be used in our further study for the structural analysis in Section 4. Therefore, only these two structural wall systems are briefly presented in the following subsections.

Massive-panel construction (CLT)

A solid timber structural system is a prefabricated massive panel timber construction where the load-bearing wall and floor elements are produced as cross-laminated planar structural elements. The main benefit of this prefabricated cross-laminated structural (CLT) system is in the perpendicular orientation of timber boards to avoid anisotropy of wood as a raw natural material. The whole production process is schematically presented in Fig. 3.

Another important advantage of CLT system over the LTF system is that its horizontal load-bearing capacity and stiffness are significantly higher, and thus such structural wall elements are mainly placed in lower storeys of a prefabricated timber building, where the internal forces due to horizontal load impacts (wind, earthquake) are the highest (Fig. 1). In the case of hybrid structural systems CLT elements are primarily placed at the envelope of the building, where the asymmetrical floor plan and the resulting distortion on the individual floors results in significantly different loads on the load-bearing wall elements due to the action of horizontal loads. In this case, the additional distortion loads are highest at the envelope of the building and lowest at the load-bearing wall elements closest to the floor shear centre [2, 16].

Fig. 3 Production process of CLT structural wall elements [16]

Light timber-framed construction (LTF); frame-panel construction

Light Timber-Framed wall elements are subdivided into two different types of technological prefabrication (Stud system – non-prefabricated and the Frame-Panel system – prefabricated). In our future implementations, we will limit ourselves to prefabricated Frame-Panel system only. The Frame-Panel system originates from the Scandinavian-American construction methods, i.e. balloon-frame and platform-frame construction types (Fig. 2), whose assembly takes place on-site. The advantages of the Frame-Panel construction system over the above-mentioned traditional timber-frame construction systems were first noticed at the beginning of the 1980s and made a significant contribution to the development of such prefabricated timber construction [16].

The load-bearing wall element consists of a timber frame, usually made up of three posts and an upper and lower beam. The upper beam transfers the vertical loads to the lower columns, which in turn transfer the vertical loads to the lower support members. The sheathing boards are attached to the timber frame by means of fasteners (staples, nails) and its tensile diagonal is of the utmost importance to transfer the loads due to the action of horizontal loads (wind, earthquake). Due to the typical dimensions of the prefabricated sheathing boards, the spacing between the columns is usually 600 mm to 625 mm. In practice, however, there are two different possible technological versions of this wall system; Single-panel (Fig. 4a) and Macro-panel system (Fig. 4b). In a statical view the Macro-panel wall assembly is considered as a sum of the contribution of all load-bearing single-panel wall elements [4, 17].

Although CLT and LTF structural systems are quite similar in terms of technology, there are significant differences in terms of construction and building-physical aspects. Thus, the CLT system is more structurally load-bearing to the effects of vertical and horizontal loads, while the LTF system shows better characteristics in terms of better thermal insulation performance for the same thickness of both wall elements. Comparison of these characteristics is widely analysed in [17].

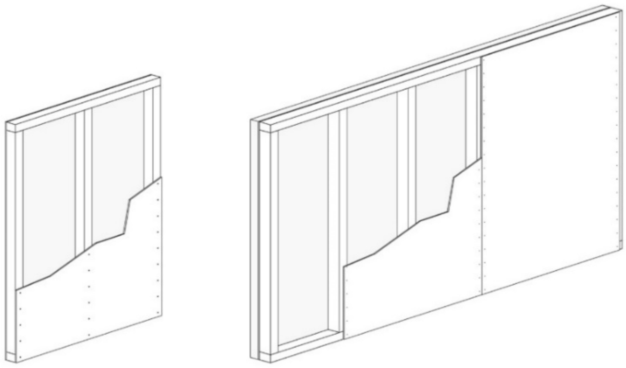


Fig. 4 a) Single-panel and b) Macro-panel prefabricated LTF wall element [16]

2.2 Load-bearing timber double-skin façade (DSF) elements

As mentioned in the introductory chapter, there is a strong tendency in contemporary timber construction to incorporate an increased proportion of glazed surfaces, both glass openings (windows, doors) and fixed glazing, which can allow increased solar radiation and better natural lighting through their transparent surfaces [4]. Usually, the asymmetric floor-plan position of such transparent facade elements and the resulting distortion on the individual floors of the building, particularly due to the action of seismic loads, has shown an increasing need to develop appropriately load-bearing timber-glass LTF wall elements that can significantly reduce these torsional effects. Thus, the so-called single-skin façade elements (SSF) were first developed as load-bearing elements. In such timber-glass wall elements a classical sheathing board (OSB or FPB) in LTF wall element presented in subsection 2.1.2. is replaced with a single glass pane which is rigidly bonded to the timber frame (Fig. 5a). The load transformation mechanism thus include a shear transmission over the glass-timber frame bonding line and the resistance in the tensile diagonal of the glass pane, as it is schematically presented in Fig. 5a. However, during many experiemental and numerical studies it was demonstrated that such elements do not prove sufficiently increased racking load-bearing capacity and in particular, do not demonstrate an importance increase in racking stiffness to improve the horizontal stiffness of the whole building [5-11]. Consequently, double-skin facade (DSF) elements were further developed in a sense to additionally improve especially the in-plane stiffness of the load-bearing transparent timber-glass wall elements.

In a case of DSF elements an additional glazing pane is added. It is important to point out that the thermal-insulating three-layered glazing is placed on the internal side of the façade element and a single-layer non-insulating glazing on the external side, as schematically presented in Fig. 5b. For exterior glazing usually a laminated heat strengthened glass is prescribed, while two- or three-layered thermal insulating glazing on inner side consists of two annealed glass panes and a safety laminated heat strengthened glass for safety reasons and for thermal insulation. The width of the cavity between the both glass panes can vary from 200 mm to even more than 2 m and can importantly influence on the U-value of a such DSF element. The frame structure can be made of steel, aluminium, plastic or timber material. However, respecting ecological impacts only the case with the timber frame will be further studied in our case in the structural analysis.

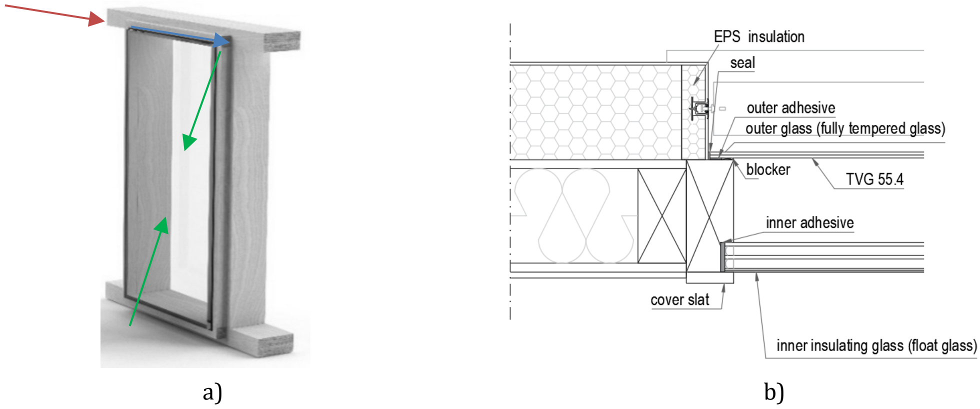


Fig. 5 a) Horizontal force distribution in SSF element; shear (in blue), tensile diagonal (in green); b) Composition of DSF load-bearing wall element [14]

Recently, many studies have analysed the thermal and acoustic performance of DSF elements [19-21]. There are also some important studies analysing the ecological impacts with various frame material [22], but almost none of them have analysed their structural behaviour, especially in terms of determining their racking resistance. All such DSF elements have been considered as in-plane non-resisting and of course in this sense also not implemented in any standards yet [18]. The numerical study in [22] is focused on the vertical load impact but does not address any racking resistance range. In a sense to study the racking behaviour of DSF elements wide experimental research was done in [12] finally resulting in European patent application product in [23]. Among that, a parametrical numerical study analysing some of the most important parameters influencing the racking resistance of DSF elements was done in [13]. Findings of this study will be directly implemented in our study of the 6-storey building in Section 4.

3. Mathematical modelling of load-bearing timber wall elements in prefabricated structures

A multi-storey prefabricated Light Timber-Framed (LTF) and cross-laminated timber (CLT) load-bearing wall elements can be effectively modelled using fictive diagonals for each lateral load-bearing wall element, as shown in Fig. 6. This approach simplifies the structural analysis of complex multi-storey timber buildings and requires significantly less computational time in comparison with all other possible approaches.

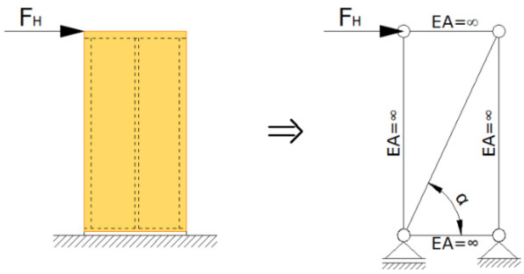


Fig. 6 Schematic presentation of the fictive diagonal model [14]

For Light Timber-Framed (LTF) walls, conventional sheathing material such as OSB or FPB are standard, but can be replaced with glass panes for single-skin (SSF) or double-skin (DSF) façade configurations. Walls with openings, such as windows or doors which are not stiff connected to the timber frame, are modelled without any diagonals and considered as non-load-bearing, as shown in Fig. 7b. The primary method for analysing these structures involves a calculation model with fictive diagonals, which simplifies the estimation of the horizontal stiffness and reduces the computational demands. The diameter of each fictive diagonal d_{fic} is calculated to ensure that the horizontal displacement of the modelled LTF wall matches that of the actual LTF wall, using the following formula:

$$d_{fic} = \sqrt{\frac{4 \cdot R \cdot L}{(\cos\alpha)^2 \cdot \pi \cdot E}} \tag{1}$$

where L is the length of the fictive diagonal, E is the elastic modulus of the fictive diagonal (E = 210 GPa if a steel bar is used for the diagonal in a calculation), and α is the inclination angle of the fictive diagonal. For LTF walls with OSB or FPB sheathing boards which are by mechanical fasteners connected to the timber frame, the racking stiffness R can be determined semi-analytically using Eq. 2. In this case, the γ-method prescribed by Eurocode 5 [18] can be used, taking into account the significant flexibility between the sheathing material and the timber frame elements. The diameter of the fictive diagonal element (Eq. 1) can be determined very rapidly in a semi-analytical form based on the effective bending stiffness $(EI)_{eff}$ calculated using the simple-beam theory and γ-method in a form of Eq. 2 by respecting the Eurocode 5 [17] expressions for the γ_i coefficient:

$$(EI)_{eff} = E_b I_b + E_t I_t = E_b \cdot \frac{n_b \cdot t \cdot b^3}{12} + E_t \cdot \left(\frac{2 \cdot a^3 \cdot c}{12} + \frac{d^3 \cdot c}{12} + 2 \cdot \gamma_i \cdot A_t \cdot z^2\right) \tag{2}$$

If the horizontal force F_H is acting at the top of the LTF wall element (Fig. 6) with the height H and the flexibility of the rocking and bottom real deformation are in this case both neglected (the both supports are for this study considered as rigid), the total flexibility of the wall element D is the sum of the in-line bending flexibility D_1 and the shear flexibility D_2 in the form of:

$$D = D_1 + D_2 = \frac{H^3}{3\,(EI)_{eff}} + \frac{H}{(GA)_{eff}} \tag{3}$$

The racking stiffness is then finally calculated in the form of:

$$R = \frac{1}{D} \tag{4}$$

For cross-laminated timber (CLT) wall elements, the racking stiffness R can be numerically obtained using special software program such as Calculatis [24], with the diameter of the fictive diagonal d_{fic} subsequently calculated for use in Eq. 1.

On the other hand, in the case of prefabricated DSF elements, the bonding line is fixed with a continuously distributed adhesive, and there are no mechanical fasteners that are point-connected to the timber frame. Therefore, respecting the Eurocode 5 [17], the γ-method cannot be adopted for the calculation for $(EI)_{eff}$ for LTF elements and consequent racking stiffness R at all and the calculation process for determining the fictive diagonal diameter d_{fic} is in this case much more complex and time consuming. In this case determination for R usually requires data from experimental studies [12] or at least a special spring model results using the finite element method (FEM) to calculate first the horizontal displacement under the acting horizontal point load F_H at the top of the wall element (Fig. 6). Crucial point in such FEM modelling is the approximation of sliding in the bonding line between both glass panes and the timber frame. This effect can be modelled by using two elastic springs in perpendicular directions (K_1 and K_2 respectively) in the form of:

$$K_1 = \frac{E_a \cdot A_a}{t_a} = \frac{E_a \cdot (w_a \cdot l_a)}{t_a} \qquad K_2 = \frac{G_a \cdot A_a}{t_a} = \frac{G_a \cdot (w_a \cdot l_a)}{t_a} \tag{5}$$

where E_a and G_a represent the modulus of elasticity and shear modulus of the adhesive, respectively, while t_a and w_a denote the thickness and width of the adhesive, respectively. The bonding length l_a serves as a parameter equal to the distance between selected springs. The mathematical modeling procedure with these springs, extensively detailed elsewhere, facilitates the determination of the racking stiffness R of load-bearing DSF elements, which can further be utilized to calculate the fictive diagonal diameter d_{fic} using appropriate equations. The whole calculation procedure is already fully described in [7, 14]. Once the horizontal displacement $u_H$under an acting force F_H is calculated the racking stiffness is finally calculated in the form of:

$$R = \frac{F_H}{u_H} \tag{6}$$

It is important to point out that this FEM calculation procedure allows for the determination of the DSF racking stiffness RRR in Eq. (1) without the need for costly and time-consuming experimental tests. This applies to DSF wall elements of arbitrary dimensions, glass pane thicknesses, and adhesive types and thicknesses.

4. Numerical study of a six-storey prefabricated timber building

4.1 Building design

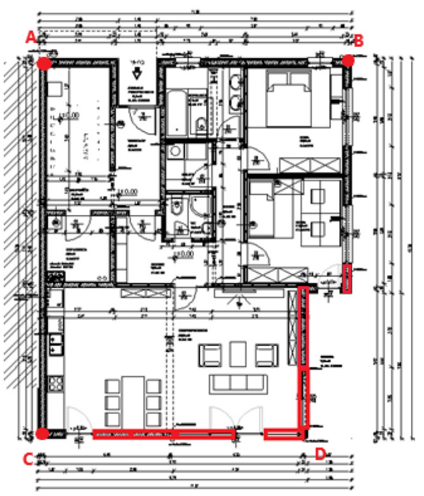
This study examines a six-storey prefabricated building specifically selected to evaluate the effects of installing additional double-skin façade (DSF) elements to increase lateral load resistance and stiffness and ensure compliance with Eurocode 5 [18] and Eurocode 8 [25] structural requirements. However, Eurocode 8 [25] does not specifically address any earthquake resistant DSF configurations. Therefore, two different calculation cases for two different structural systems (hybrid and non-hybrid) are performed:

- Considering DSF wall elements as non-load-bearing to evaluate their effect on horizontal loads according to current Eurocode 8 [25] requirements;
- DSF wall elements are treated as in-plane load-bearing to evaluate their effect on increased horizontal load resistance of the whole building.

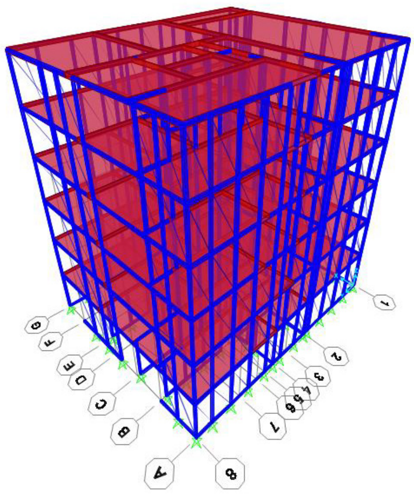
A mathematical analysis of a six-storey prefabricated timber building with a maximal height H of 15 m is carried out, with particular emphasis on the horizontal stiffness and the natural frequencies of oscillation. The ground floor plan is shown in Fig. 7a and the building complies with the height requirements of Eurocode 8 [25]. The building envelope is bounded by load-bearing walls from the first floor to the top floor.

In the first studied case (Case 1), all internal wall elements consist of load-bearing walls in Light Timber-Framed (LTF) construction accompanied by traditional fibreboard (FPB) sheathing, identified on the floor plans by black filler. In this hybrid structural wall system solution all external walls on the building envelope are constructed from higher resistant cross-laminated timber (CLT) panels and, like the interior walls, are made from C24 class timber.

In the second analysed case (Case 2) in a sense to increase the overall racking stiffness of the whole building and to ensure the prescribed Eurocode conditions for maximal horizontal displacements [18, 25], both the internal and external wall elements are composed of CLT timber components only and the structural system is thus non-hybrid. There are many advantages and disadvantages of CLT and LTF structural systems according to the structural and non-structural facts which are deeply studied in [17] where also many benefits of hybrid structural solutions are discussed.



a)



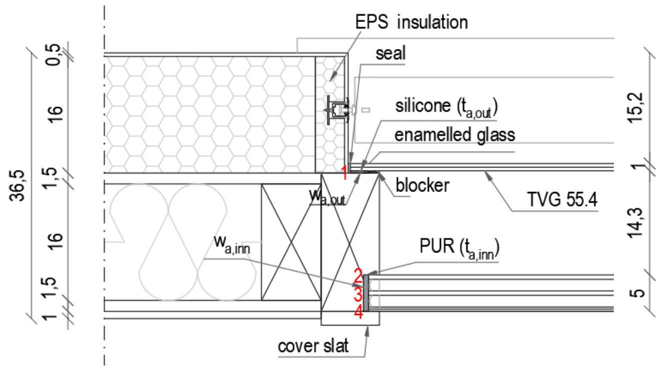
b)

Fig. 7 a) Floor-plan of the building; b) Computational model of the structure made by SAP 2000 software

In both cases the transparent DSF wall sections are marked with red markings (Fig. 7a). Other transparent areas, which represent windows and doors, are shown with white infill and are treated as non-load-bearing elements in the structural analysis in performed numerical cases. These components do not contribute to the load-bearing capacity of the building and are excluded from the structural integrity calculations at all.

The computational model shown in Fig. 7b was developed using the structural analysis and dynamics software SAP 2000. Although the building meets the plan correctness required by the Eurocode 8 [25], a 3D structural model was used. The analysis includes all relevant load-bearing wall elements, which are represented by fictive diagonals. Their effective cross-sectional areas are calculated based on Eqs. 1-3 for Light timber-framed (LTF) walls with FPB sheathing and Eqs. 2 and 4 for load-bearing DSF walls. For CLT walls, the racking stiffness R is determined using the Calculatis program [24] and further used in Eq. 1 to calculate the effective cross-sectional diameter d_{fic}.

Fig. 8 shows a cross-sectional view of a DSF wall element with different adhesives. Two types of adhesives are used: Polyurethane Ködiglaze P for the inner triple thermal insulation glazing and Silicone Ködiglaze S for the outer single glazing. The timber frame of the construction consists of glulam (GL24h) in accordance with classification EN 1194 [26]. The internal glass pane consists of thermal-insulating three-layered glazing while the outer glass pane is single-layer and is made of toughened laminated glass.



1. Laminated outer glass (5+5 mm)
2. Float glass (6 mm)
3. Float glass (6 mm)
4. Laminated glass (4+4 mm)

Fig. 8 Schematic presentation of a DSF load-bearing structural wall element

The input data for two types of adhesives and the material properties of the glass and timber elements are listed in Tables 1 and 2, respectively.

Table 1 Material properties of the adhesives [27, 28]

	Poisson's ratio ν	Shear modulus G_s (MPa)	Elastic modulus E_0 (MPa)	Decay time t_d (s)	Decay constant β
Silicone (Ködiglaze S)	0.5	0.351	1.053	100	0.0026
Polyurethane (Ködiglaze P)	0.49	0.454	1.354	290	0.0016

Table 2 Material properties of the timber [26] and glass components [29, 30]

	Timber frame GL24h [26]	Float glass [29]	Thermally toughened glass [30]
Standard	EN 1194	EN 12150	EN 12150
E (Mpa)	II 11,600 ⊥ 720	70,000	70,000
ν (-)	II 0.25 ⊥ 0.45	0.23	0.23
G (MPa)	II 720 ⊥ 35	0.45	0.45
f_t (MPa)	II 14 ⊥ 0.5	45	120
f_c (MPa)	II 14 ⊥ 0.5	500	500
ρ (kg/m^3)	380	2,500	2,500

4.2 Numerical analysis

The polyurethane adhesive used in the calculations for fixing the internal glazing has a thickness of $t_{a,inn}$ = 7 mm and a width of $w_{a,inn}$ = 28 mm, parameters that match those of the experimentally tested DSF specimens [12]. The influence of the polyurethane adhesive with additional parametrically chosen values of thickness $t_{a,inn}$ = 3, 5, 7 and 9 mm) on a racking stiffness of a single DSF wall element is numerically studied in [13]. Additionally, a huge experimental study using polyurethane and epoxy adhesive with an emphasis on the comparison of experimental results with SSF elements is given in [12].

Considering the extensive load-bearing structure of the whole building, the application of the spring model calculation is computationally prohibitive. Therefore, we use the mathematical model with fictive diagonals as described in Section 3. The cross-section diameter d_{fic} of the fictive diagonal elements is determined using Eqs. 2-6, with the calculated values varying depending on the type of a prefabricated timber wall element (LTF, CLT, DSF).

Table 3 presents the calculated fictive diagonal diameters and the racking stiffness for each type of resistant wall element: FPB sheathing boards for internal timber wall elements, load-bearing CLT elements for external/internal wall elements and load-bearing DSF elements. Both presented values for the racking stiffness R could be used for the DSF elements: 909 N/mm from the comprehensive experimental study [12] or 857 N/mm from the elastic FEM spring model [13]. However, to simplify the whole procedure only the value of 857 N/mm is further used for the numerical analysis of the entire six-storey building.

The calculated cross-sectional values of the fictive diagonal d_{fic} show that the racking stiffness of CLT wall elements is significantly higher than that of Light Timber-Framed (LTF) wall elements with conventional fibre-plaster sheathing boards (FPB) and also significantly higher compared to the resistant DSF elements. The problem of the relatively low in-plane stiffness of DSF elements was comprehensively analysed and discussed in [13] both based on experimental results and in a subsequent parametric numerical study using an elastic spring FEM model.

Table 3 Diameter of the fictive diagonals and load-bearing capacities of the wall elements

Load-bearing wall elements	Racking stiffness R of the resisting wall elements (N/mm)	Diameter of the fictive diagonal d_{fic} (mm)
DSF (experimental)	909	8.78
DSF (spring model)	857	8.52
CLT external wall	5602	21.79
LTF internal wall	3425	17.04

4.3 Numerical results and discussion

The oscillation times (first three modes) of the six-storey building for the two analysed cases load-bearing and non-load-bearing DSF elements are first calculated using a 3D FEM model (Fig. 7b) with fictive diagonals and SAP 2000 software. In the first case, the interior of the building consists of load-bearing walls in Light Timber-Framed (LTF) construction, supplemented by traditional fibreboard (FPB) sheathing. The exterior walls are made of cross-laminated timber (CLT) panels and like the interior walls are made from C24 timber. In the second case in a sense to enlarge the overall racking stiffness of the building, both the internal and external wall elements are made of CLT timber exclusively. The results for the calculated first three natural oscillation modes (T_1, T_2, T_3) for both cases, taking into account the stiffness contribution of non-load-bearing and load-bearing DSF elements, are shown in Table 4.

Table 4 Oscillation times (T_1, T_2, T_3) of the six-storey building considering both cases for load-bearing and non-load-bearing DSF wall elements

DSF element	Non-load-bearing DSF elements		Load-bearing DSF elements	
Oscillation mode	Oscillation times T (s)		Oscillation times T (s)	
	Case 1 (CLT+LTF)	Case 2 (all in CLT)	Case 1 (CLT+LTF)	Case 2 (all in CLT)
1. (T_1)	0.934	0.840	0.839	0.765
2. (T_2)	0.721	0.672	0.689	0.645
3. (T_3)	0.516	0.506	0.507	0.498

As expected, the oscillation times are higher when non-load-bearing DSF elements are considered, which is due to the lower overall racking stiffness of the structure, while the mass remains unchanged. Also, in the first case (hybrid CLT+LTF), where the interior of the building consists of load-bearing walls of Light timber-framed construction (LTF) and the exterior walls of cross-laminated timber panels (CLT), the oscillation times are higher than in the second case, where the internal and external wall elements consist of CLT timber. Among that, it can be observed, that the oscillation times are significantly reduced in both cases by considering the additional stiffness of the DSF elements. This is particularly evident for the first oscillation time T_1 and least evident for the third oscillation time T_3. The decrease in T_1 for Case 1 is 11.13 % and for Case 2 8.23 %. It is a quite logical because the overall stiffness is higher in the case of non-hybrid CLT structural wall system and therefore an additional contribution of DSF is less evident.

Tables 5 and 6 present the calculated horizontal racking stiffnesses R and displacements of the structure in the two global orthogonal directions of seismic action (X and Y directions) at selected control points (A-D, see Fig. 9). For the calculation of the displacements, a rather large seismic intensity with a_g = 0.225·g is deliberately chosen. Numerical analysis under a higher random excitation of a_g = 0.30· g and a_g = 0.40·g performed on one and two-storey timber box-house models previously experimentally tested in [8] with SSF elements is additionally presented in [11].

Both cases with load-bearing and non-load-bearing DSF elements are considered in these calculations. According to [25], the allowed value of horizontal displacements in a multi-storey building is H/500, which corresponds to 30 mm in our case.

Table 5 Racking stiffnesses R and displacements u of the corner points on the top storey of the six-storey building for Case 1 (hybrid CLT+LTF)

	DSF Element	Non-load-bearing DSF elements		Load-bearing DSF elements	
	Earthquake	Direction X	Direction Y	Direction X	Direction Y
Point	Displacement (mm)				
A	u_x	19.27	21.51	19.56	20.46
	u_y	7.59	15.51	6.20	15.05
	u_R	20.71	26.52	20.52	25.40
B	u_x	19.26	21.50	19.56	20.45
	u_y	28.12	26.81	22.61	26.87
	u_R	34.08	34.37	29.90	33.77
C	u_x	36.32	22.38	33.58	17.97
	u_y	7.60	15.51	6.21	15.06
	u_R	37.11	27.23	34.15	23.45
D	u_x	36.33	22.39	33.58	17.97
	u_y	28.15	26.85	22.63	26.89
	u_R	45.96	34.96	40.49	32.34
	R (N/mm)	6639	9209	8168	10169
EC 8 requirement [25] u_{max} = H/500		30.00	30.00	30.00	30.00

It is presented again that the increase in overall structure racking stiffness (R) by using additional DSF elements as load-bearing is very influent. For instance, in the X-direction this increase is of 23.03 % and in the Y-direction 10.42 %. Again, respecting the floor-plan design in Fig. 7a, it is logical, because practically all façade elements in south (X) direction are made from DSF wall components. On the other hand, in the east orientation (Y) direction the percentage of DSF elements is essentially lower.

In this case of completely non-hybrid and more rigid CLT structure the increase of overall racking stiffness R is of 17.26 % in X-direction and 16.27 % in the Y-direction. So, compared to Case 1, the increase in the X-direction is significantly smaller, but even slightly larger in the Y-direction, which is due to the fact that more load-bearing DSF elements are placed on the south façade (X-direction) than on the east façade (Y-direction). Additionally, the observed maximal horizontal displacements exceed the limits prescribed by Eurocode 8 [25] by approximately 53 % (Case 1) and 33 % (Case 2) if the DSF elements are considered as non-resisting. However, by considering DSF elements as resisting elements, the exceeded values for horizontal displacements essentially decrease and are higher only for 35 % (Case 1) and 20 % (Case 2). These values are highlighted in red in the table. Of course, if the selected seismic excitation would be a little lower (less than $0.15\cdot g$), these Eurocode conditions could easily be met for the case of a 6-storey building.

Table 6 Racking stiffnesses R and displacements u of the corner points on the top storey of the six-storey building for Case 2 (all in CLT)

	DSF Element	Non-load-bearing DSF elements		Load-bearing DSF elements	
	Earthquake	Direction X	Direction Y	Direction X	Direction Y
Point	Displacement (mm)				
A	u_x	20.21	19.48	19.35	18.79
	u_y	7.36	15.20	5.99	14.57
	u_R	21.51	24.71	20.26	23.78
B	u_x	20.21	19.48	19.35	18.78
	u_y	26.08	25.22	22.17	24.56
	u_R	32.99	31.87	29.43	30.92
C	u_x	30.01	23.00	28.02	19.26
	u_y	7.36	15.20	5.99	14.57
	u_R	30.90	27.57	28.65	24.15
D	u_x	30.01	23.00	28.02	19.26
	u_y	26.11	25.25	22.18	24.59
	u_R	39.78	34.15	35.74	31.23
	R (N/mm)	8623	9865	10111	11470
EC 8 requirement [25] $u_{max} = H/500$		30.00	30.00	30.00	30.00

As this is a 6-storey timber structure, where usually considerable problems with the prescribed Eurocode Serviceability limit state conditions [17, 25] in ensuring for maximum horizontal displacements ($H/500$) already exist, it is also interesting to compare the contribution of the load-bearing DSF elements compared to the contribution of the use of CLT wall elements also for the internal wall elements. Comparing the values for R in the X- and Y-direction between Tables 6 and 5, it can be seen that the use of load-bearing DSF elements in the hybrid structural system (CLT+LTF; Case 1) results in essentially very similar stiffnesses (R_x = 8168 N/mm and R_y = 10169 N/mm) as in the CLT-only structural system (Case 2), and where all DSF elements are non-load-bearing (R_x = 8623 N/mm and R_y = 9865 N/mm).

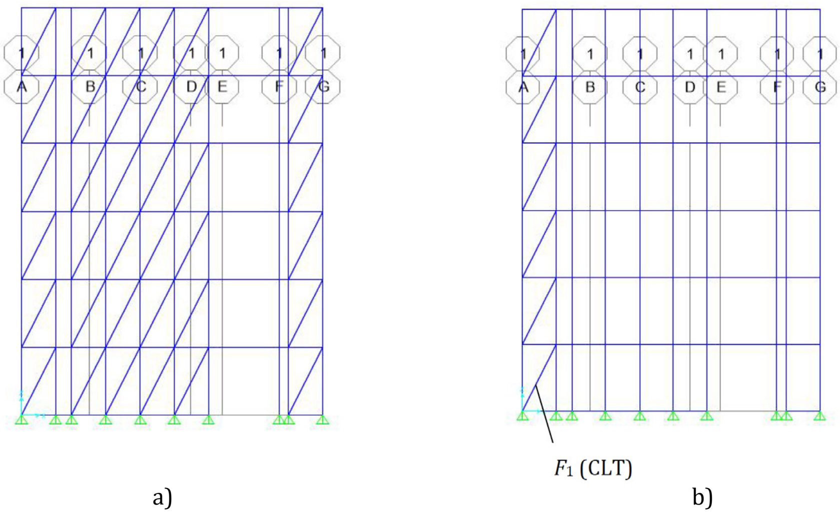


Fig. 9. View along Axis 1 of the six-storey wall model using: a) load-bearing DSF elements and b) non-load-bearing DSF elements

Finally, at the end of the study, the horizontal force acting on the CLT corner wall element 1 along the axis 1 (F_1 in Fig. 9b) is specifically monitored to evaluate a possible reduction of the torsional effects caused by the asymmetric plan due to the positioning of the transparent DSF elements. It is assumed that the use of additional DSF elements as load-bearing components, especially on the south façade, will help to reduce the force acting on the primary load-bearing CLT corner element. Tables 7 and 8 show the calculated horizontal forces for Cases 1 and 2 in both directions (F_x and F_y) acting on this CLT wall element due to seismic actions in the two global orthogonal directions (X and Y). In these tables, the results are compared for both configurations with non-load-bearing and with load-bearing DSF elements. The resulting diagonal force F_R acting on this CLT element is also calculated.

Table 7 Horizontal forces in the exterior walls of the building (Case 1 – hybrid CLT+LTF)

DSF Element		Non-Load-Bearing DSF Elements		Load-Bearing DSF Elements	
Earthquake		Direction X	Direction Y	Direction X	Direction Y
Axis	Force (kN)				
1	F_x	29.25	18.37	29.78	16.02
	F_y	89.56	63.05	96.38	62.95
	F_R	94.22	65.67	100.88	64.96
8	F_x	16.56	19.17	15.87	19.29
	F_y	66.40	78.94	59.81	80.81
	F_R	68.43	81.23	61.88	83.08
A	F_x	8.02	16.76	6.68	16.25
	F_y	89.56	63.05	96.38	62.95
	F_R	89.92	65.24	96.61	65.01
G	F_x	0.09	0.08	0.08	0.08
	F_y	4.15	2.90	20.93	11.68
	F_R	4.15	2.90	20.93	11.68
Axis 1:	F_1	29.25	63.05	29.78	62.95

Table 8 Horizontal forces in the exterior walls of the building (Case 2 – all in CLT)

DSF Element		Non-Load-Bearing DSF Elements		Load-Bearing DSF Elements	
Earthquake		Direction X	Direction Y	Direction X	Direction Y
Axis	Force (kN)				
1	F_x	23.52	18.35	24.08	16.76
	F_y	71.98	62.23	96.38	62.10
	F_R	75.73	65.84	78.98	64.34
8	F_x	17.41	16.63	16.18	16.95
	F_y	72.54	71.84	65.09	73.78
	F_R	74.60	73.74	67.07	75.70
A	F_x	7.39	16.04	6.03	15.30
	F_y	71.68	62.23	78.98	62.10
	F_R	72.06	64.26	79.21	63.96
G	F_x	0.07	0.07	0.07	0.07
	F_y	3.98	3.01	18.11	12.48
	F_R	3.98	3.01	18.11	12.48
Axis 1:	F_1	23.52	62.23	24.08	62.10

The results presented show that in the case of load-bearing DSF elements, the force F_1 in the X-direction was reduced by 1.8 % (Case 1) and by 2.3 % (Case 2), while in the Y-direction the force remained approximately the same in both cases. This is quite logical, as Table 3 shows that the stiffness of the CLT beams is significantly higher than of the LTF wall elements. This means that the contribution of considering DSF as load-bearing elements in LTF structures is much more important. In addition, the reduction of the acting force in the X-direction F_x on the CLT member is lower than in the similar study [14], in which a three-storey building with the same floor plan built exclusively in a Lightweight Timber-Framed (LTF) system was analysed.

5. Conclusion

The numerical study carried out on the selected 6-storey prefabricated timber structure clearly demonstrated the importance of considering the previously developed and pa-tended timber DSF elements [23] as additional bracing envelope wall components of the structure. In the analysis, we have deliberately chosen two computational cases, where first the importance of the DSF load-bearing elements has been analysed in the case of a hybrid structural system (CLT+LTF), and secondly in the case of a generally more rigid non-hybrid full CLT system. Also, all DSF elements were deliberately placed rather asymmetrically around the building envelope, entirely on the south façade (in the X-direction) and partially also on the east façade (in the Y-direction), as is also the norm in contemporary multi-storey timber buildings.

In both cases, the results of the studies carried out showed a significant increase in the horizontal stiffness of the whole building when DSF elements are considered as load-bearing. In Case 1, this increase was 23 % in the X-direction and 10.4 % in the Y-direction, while in Case 2 it was 17.3 % and 16.3 % respectively, i.e. much more symmetric. Consequently, the increased horizontal stiffness of the whole building also makes it much easier to meet the prescribed Eurocode conditions for maximum displacements in the case of using load-bearing DSF elements.

However, as already analysed in detail in [17], hybrid structural systems are usually the most optimal solution in multi-storey prefabricated timber buildings when several different aspects, both structural, structural-physical and environmental, are considered. Particularly important in this respect is the combination of Light Timber-Framed (LTF) and Solid-timber system in the form of Cross-laminated Timber (CLT), which was basically considered in the design of our structure in Case 1. With Case 2, where all the internal load-bearing wall elements constructed in LTF system were replaced by the more rigid CLT, we only followed the basic design objective of increasing the horizontal stiffness of the whole building envelope, but not in some other also very important respects. However, the analysis carried out has shown that a very similar increase in the horizontal stiffness of the whole building can be achieved in fact by considering the pa-tented DSF elements as additional load-bearing elements. In this case, of course, there is further no need to design the building exclusively in CLT structural wall system, which may be inferior in some respects to a hybrid (CLT+LTF) design.

The results of the study highlight the importance of carefully considering the load-bearing DSF elements in seismic design, as their incorporation can have a significant impact on the overall behaviour of the multi-storey timber structure through their influence on the displacements and stiffness properties. In a sense of European standards, it exist currently only some guidance for European structural design of timber-glass components [31]. Taking these considerations into account is crucial to ensure compliance with seismic design codes [25] and to improve the structural resilience of multi-storey prefabricated timber buildings in a sense to be further implemented also in Eurocode 8 standard.

Acknowledgement

Funding for this research was provided by the Slovenian Research Agency, National research programme P2-0129.

References

[1] Li, J., Rismanchi, B., Ngo, T. (2019). Feasibility study to estimate the environmental benefits of utilising timber to construct high-rise buildings in Australia, *Building and Environment*, Vol. 147, 108-120, doi: 10.1016/j.buildenv.2018.09.052.

[2] Zegarac Leskovar, V., Premrov, M. (2021). A review of architectural and structural design typologies of multi-storey timber buildings in Europe, *Forests*, Vol. 12, No. 6, Article No. 757, doi: 10.3390/f12060757.

[3] Salvadori, V. (2021). *Multi-storey timber-based buildings: An international survey of case studies with five or more storeys over the last twenty years*, Dissertation, Technische Universitat Wien (TU Wien), Vienna, Austria.

[4] Zegarac Leskovar, V., Premrov, M. (2013). *Energy-efficient timber-glass houses*, Springer, London, United Kingdom, doi: 10.1007/978-1-4471-5511-9.

[5] Neubauer, G., Schober, K.P. (2008). *Holz-Glas-Verbundkonstruktionen*, Weiterentwicklung und Herstellung von Holz-Glas-Verbundkonstruktionen durch statisch wirksames Verkleben von Holz und Glas zum Praxiseinsatz im Holzhausbau (Impulsprojekt V2 des Kind Holztechnologie), Endbericht, Holzforschung Austria, Vienna, Austria.

[6] Cruz, P., Pequeno, J. (2008). Timber-glass composite structural panels: experimental studies & architectural applications, *Conference on Architectural and Structural Applications of Glass*, Delft University of Technology, Faculty of Architecture, Delf, Netherlands.

[7] Hochhauser, W. (2011). *A contribution to the calculation and sizing of glued and embedded timber-glass composite panes*, Doctoral Thesis, Vienna University of Technology, Faculty of Civil Engineering, Austria.

[8] Premrov, M., Serrano, E., Winter, W., Fadai, A., Nicklisch, F., Dujic, B., Sustersic, I., Brank, B., Strukelj, A., Drzecnik, M., Buyuktaskin, H.A., Erol, G., Ber, B. (2015). Workshop report, *WP 6: Testing on life-size specimen components:*

shear walls, beams and columns including long-term behaviour, WoodWisdom-net, research project, load bearing timber-glass-composites, 2012-2014, 15.
[9] Strukelj, A., Ber, B., Premrov, M. (2015). Racking resistance of timber-glass wall elements using different types of adhesives, *Construction and Building Materials*, Vol. 93, 130-143, doi: 10.1016/j.conbuildmat.2015.05.112.
[10] Ber, B., Finzgar, G., Premrov, M., Strukelj, A. (2019). On parameters affecting the racking stiffness of timber-glass walls, *Glass Structures & Engineering*, No. 4, 69-82, doi: 10.1007/s40940-018-0086-5.
[11] Premrov, M., Ber, B., Kozem Silih, E. (2021). Study of load-bearing timber-wall elements using experimental testing and mathematical modelling, *Advances in Production Engineering & Management*, Vol. 6, No. 1, 67-81, doi: 10.14743/apem2021.1.385.
[12] Premrov, M., Strukelj, A., Ber, B. (2023). Experimental study on racking resistance of timber double-skin façade elements, *Engineering Structures,* Vol. 293, Article No. 116712, doi: 10.1016/j.engstruct.2023.116712.
[13] Kozem Silih, E., Premrov, M. (2022). Numerical study of racking resistance of timber-made double-skin façade elements, *Advances in Production Engineering & Management*, Vol. 17, No. 2, 231-242, doi: 10.14743/apem2022.2.433.
[14] Premrov, M., Kozem Silih, E. (2023). Numerical analysis of the racking behaviour of multi-storey timber-framed buildings considering load-bearing function of double-skin façade elements, *Sustainability*, Vol. 15, No. 8, Article No. 6379, doi: 10.3390/su15086379.
[15] Kolb, J. (2008). *Systems in timber engineering*, Birkhäuser Verlag, Basel, Switzerland.
[16] Premrov, M. (2020). Structural design of sustainable timber buildings, In: Skalicky Klemenčič, V., Žegarac Leskovar, V, Žigart, M. (eds.), *An interdisciplinary approach towards academic education on sustainable building design*, Nova Science Publishers, Hauppauge, USA, 129-164.
[17] Premrov, M., Zegarac Leskovar, V. (2023). Innovative structural systems for timber buildings: A comprehensive review of contemporary solutions, *Buildings,* Vol. 13, No. 7, Article No. 1820, doi: 10.3390/buildings13071820.
[18] European Committee for Standardization. EN 1995-1-1; Eurocode 5: Design of timber structures. General. Common rules and rules for buildings, from *https://www.phd.eng.br/wp-content/uploads/2015/12/en.1995.1.1.2004.pdf*, accessed February 12, 2024.
[19] Chan, A.L.S., Chow, T.T., Fong, K.F., Lin, Z. (2009). Investigation on energy performance of double skin façade in Hong Kong, *Energy and Buildings*, Vol. 41, No. 11, 1135-1142, doi: 10.1016/j.enbuild.2009.05.012.
[20] Tijjani Musa, B., Alibaba, H.Z. (2015). Evaluating the use of double skin facade systems for sustainable development, *International Journal of Recent Research in Civil and Mechanical Engineering,* Vol. 2, No. 2, 151-159.
[21] Xu, L., Ojima, T. (2007). Field experiments on natural energy utilization in a residential house with a double skin façade system, *Building and Environment,* Vol. 42, No. 5, 2014-2023, doi: 10.1016/j.buildenv.2005.07.026.
[22] Pomponi, F., D´Amico, B. (2017). Holistic study of a timber double skin façade: Whole life carbon emissions and structural optimization, *Building and Environment,* Vol. 124, 42-56, doi: 10.1016/j.buildenv.2017.07.046.
[23] Premrov, M., Zegarac Leskovar, V., Kozem Silih, E., Ber, B. (2024). *Load-bearing prefabricated double-skin façade timber wall element*: European Patent specification EP3985191 (B1), 2024-02-07 = Lasttragendes vorgefertigtes zweischaliges Fassadenholzwandelement = Élément de mur de façade en bois porteur préfabriqué à double peau. Munich: European Patent Office, 2024. 6, 2 p.
[24] Calculatis. Stora Enso, from *https://calculatis.storaenso.com*, accessed February 2, 2024.
[25] European Committee for Standardization. EN 1998-1-1: Eurocode 8: Design of structures for earthquake resistance, Part 1: General rules, seismic actions and rules for buildings, from *https://www.phd.eng.br/wp-content/uploads/2015/02/en.1998.1.2004.pdf*, accessed February 12, 2024.
[26] European Committee for Standardization. EN 1194:2003: Timber structures - glued laminated timber strength classes and determination of characteristic values, Brussels, Belgium.
[27] Kömmerling. Product information Ködiglaze P, *Special adhesive for bonding insulating glass units into the window sash*, from *https://ragasztastechnika.com/wp-content/uploads/2016/01/P_koe_ps_gb.pdf*, accessed February 12, 2024.
[28] Kömmerling. Product information Ködiglaze S, Special adhesive for structural and direct glazing, from *https://www.fgtrading.co.za/documents/sealants/kodiglazes_tds.pdf*, accessed February 12, 2024.
[29] European Committee for Standardization. EN 572-2:2004, Glass in building – Basic soda lime silicate glass products – Part 2: Float glass, from *https://standards.iteh.ai/catalog/standards/cen/1de4aa69-16e7-4297-8d3a-b38b96c73ca8/en-572-2-2004?srsltid=AfmBOook6L2_vn4os5swgjYRX1VaXt4WO9tiilyZUaZEwyCgnTiAwTs2*, accessed February 12, 2024.
[30] European Committee for Standardization. EN 12150-1:2000, Glass in building – Thermally toughened soda lime silicate safety glass – Part 1: Definition and description, from *https://standards.iteh.ai/catalog/standards/cen/91910462-d649-45cf-9b18-0b3665526d1c/en-12150-1-2000?srsltid=AfmBOoqEG8lr5fgXMyhuwi8Je2a614oXpRUvb6AaxwZHk1v02TGeAboF*, accessed February 12, 2024.
[31] Feldmann, M., Kaspar, R., Abeln, B., Gessler, A., Langosch, K., Beyer, J., Schneider, J., Schula, S., Siebert, G.; Haese, A., Wellershoff, F., Cruz, P., Belis, J., Colvin, J., Morgan, T., Ensslen, F., Eliasova, M., Šulcová, Z., Royer-Carfagni, G., Galuppi, L., Grenier, C., Hoegner, H., Kruijs, R., Louter, C., Manara, G., Neugebauer, J., Rajcic, V., Zarnic, R. (2014). *JRC scientific and policy reports, Guidance for European structural design of glass components*, Publication Office of the European Union, Luxembourg, doi: 10.2788/5523.

APEM journal
Advances in Production Engineering & Management
Volume 20 | Number 1 | March 2025 | pp 131–149
https://doi.org/10.14743/apem2025.1.532
ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

From Zero to One: A new perspective on the fuzzy front end of innovation and the Stage-Gate® model

Alešnik, P.[a,*], **Vrečko, I.**[b], **Palčič, I.**[a]

[a]University of Maribor, Faculty of Mechanical Engineering, Maribor, Slovenia
[b]University of Maribor, Faculty of Economics and Business, Maribor, Slovenia

A B S T R A C T

The Stage-Gate® model has historically provided a systematic framework for New Product Development (NPD). However, the evolving landscape of innovation necessitates continuous enhancement. This paper redefines the model's foundational structure by advocating for the recognition of the Discovery Phase as Stage 1, emphasizing its essential role in aligning initial ideation with strategic goals, streamlining processes, and enhancing NPD efforts. Using a mixed-methods approach, including a systematic literature review, synthesis of illustrative examples and secondary data and case study analysis, the research demonstrates that formalizing the Discovery Phase improves early-stage decision-making, enhances alignment between front-end exploration and downstream execution and mitigates risks by supporting more informed project development. Synthesised sectoral examples show that incorporating the Discovery Phase improves feasibility, reduces risk, and boosts efficiency. For example, simulation planning early in innovation process increased manufacturing throughput by 52 %, while early IP checks lowered infringement risk. The proposed revision boosts the Stage-Gate® model's adaptability and integration with modern methodologies such as AI, Agile, Lean Startup, Design Thinking and TRIZ. The findings highlight how this change promotes a comprehensive approach to NPD. The implications extend to practical applications and future research, offering organizations a flexible framework that meets modern market and technological demands.

A R T I C L E   I N F O

Keywords:
Stage-Gate® model;
Fuzzy front end of innovation (FFEI);
New product development (NPD);
Innovation management;
Discovery phase;
Agile;
TRIZ;
Design thinking;
Large language model (LLM);
Sustainability

**Corresponding author:*
peter@alesnik.com
(Alešnik, P.)

Article history:
Received 15 November 2024
Revised 27 February 2025
Accepted 15 March 2025

1. Introduction

The Stage-Gate® model, developed by Robert G. Cooper in the 1980s, has been a foundational framework to guide New Product Development (NPD) across a wide range of industries [1]. Over time, this model has evolved to meet the changing needs of businesses, reflecting the dynamic nature of innovation, but its fundamental structure–stages separated by decision-making gates–has remained the same [2, 3]. Traditionally, the model begins with the Discovery Phase, also known as Idea Stage, "pre-stage" or "Stage 0", often treated as an implicit, preliminary, or optional stage [1, 4]. The Stage-Gate® model has not always given the Discovery Phase the same weight as its subsequent phases, despite its critical role in idea generation and opportunity exploration [5]. Because the label 'Stage 0' can suggest a mere preparatory footnote, we refer to the Discovery Phase as Stage 1 to signal that it stands on equal footing with the later stages. This proposed renumbering is further substantiated in Section 6 and also echoes Thiel's '0 to 1' metaphor, which portrays innovation as the leap from nothing to the first spark in the innovation process—precisely what happens during the Discovery Phase [6].

The Discovery Phase is described as the "fuzzy front end", characterised by ambiguity, creativity, and the exploration of new possibilities [7-9]. It involves activities such as opportunity exploration, market analysis, and alignment with organisational goals [1, 2] and, in some cases, can function as a process on its own [10]. Despite its significance, this initial stage has been underemphasised in both practice and literature, often considered "pre-work" or a preliminary and implicit activity rather than a formal part of the process [1, 4, 5, 9]

Our research explores the evolution of the Stage-Gate® model, with a focus on the rationale for recognising the Discovery Phase as an integral part of the innovation process by establishing it as a Stage 1 rather than Stage 0. Through an analysis of historical trends in innovation models and the increasing importance of early-stage decision-making [11-13], we argue that redefining Stage 1 offers significant theoretical and practical benefits. In particular, we examine how formalising this phase aligns it more closely with the rest of the NPD process, making it an inseparable part of the innovation framework.

Furthermore, we discuss the integration of modern innovation techniques, such as open innovation and data-driven decision-making, and how these approaches fit within the redefined structure. These techniques enable organisations to respond more effectively to emerging market trends and technological advancements [7, 14-17], solidifying the Discovery Phase as an essential part of the Stage-Gate® model.

This study begins by examining the origins and development of the Stage-Gate® model, highlighting key adaptations over time. The central argument focuses on the renumbering of the Discovery Phase to Stage 1, supported by theoretical foundations, practical insights, and case studies. We then discuss modern approaches to innovation, illustrating how they align with the revised structure of the Stage-Gate® model. Finally, we outline key implications for future research and practical applications, providing a forward-looking view on NPD practices.

2. Methodology and research approach

We used a mixed-methods research design, combining a systematic literature review (SLR), synthesis of illustrative examples and secondary data and case study analysis to deepen understanding of the Stage-Gate® model and support redefining the Discovery Phase as Stage 1. This approach structured our exploration of existing literature and real-world examples, strengthening the conceptual framework developed throughout the study.

The SLR systematically gathered, synthesised, and evaluated existing literature on the Stage-Gate® model and its adaptations in the context of modern innovation methodologies. The review followed a defined search strategy, using databases such as Scopus, Web of Science, and Google Scholar, with search terms including "Stage-Gate model", "New Product Development", "Discovery Phase", "Case Study", and "Fuzzy Front End". We restricted the inclusion criteria to peer-reviewed academic papers, industry reports, and case studies published between 2000 and 2024, with most the sources falling within this period, ensuring a focus on recent advancements in innovation management. Key gaps in the literature were identified, particularly the need to redefine the Discovery Phase in light of its actual usage. These findings laid the foundation for the conceptual proposition developed in this study.

To complement the SLR, a detailed analysis of case studies and illustrative examples was incorporated to demonstrate the application of innovation methodologies within organisations. These examples, drawn from published industry cases and validated secondary sources, were synthesised to contextualise how Discovery Phase practices vary across sectors. This analysis highlights how elements of the Discovery Phase are often implicitly integrated into broader NPD frameworks, supporting the argument for formalising this phase as an integral part of the Stage-Gate® model.

The research proposes a significant adaptation, redefining the Discovery Phase as a theoretical contribution that structures early-stage ideation in NPD. While the conceptual arguments are supported by the SLR and case study analysis, further empirical research is needed to validate these propositions across various industries and company sizes. Future studies should focus on testing the effectiveness of the proposed renumbering in different industrial contexts and investigating how companies that do not explicitly define a Discovery Phase manage early-stage ideation through other stages of the Stage-Gate® model.

The proposed framework provides a foundational, conceptual model that requires further testing and refinement through empirical studies. The adaptation offers flexibility for modern innovation management, but it must be validated to retain the core efficiency of the Stage-Gate® model.

3. Systematic innovation process

In increasingly competitive markets, innovation, a critical component of modern business strategy, drives growth and success [18, 19]. The innovation process is a complex undertaking that consists of numerous internal and external processes [20, 21]. Through the NPD, the systematic innovation process aligns creativity with strategic goals and market demands [22-24].

The Stage-Gate® model is a prime example of a systematic innovation process [3]. The structured approach offers a clear framework for managing product development from concept to launch, ensuring that innovation efforts are organised, repeatable, and strategically focused [25]. By dividing the innovation process into distinct stages with decision gates, the Stage-Gate® model allows for thorough evaluation and planning at each stage, which helps to mitigate risk and improve efficiency. This systematic approach ensures that organisations maintain alignment between creative exploration and practical constraints, resulting in focused, efficient, and effective innovation [26].

For example, the pharmaceutical industry often employs a systematic innovation process for drug development, ensuring close alignment between research, development, and regulatory bodies [27]. Such alignment not only accelerates the time to market but also ensures compliance with stringent regulations [16]. The high failure rate in NPD is a significant challenge, with top performers succeeding while others face significant failure [28, 29].

To address the challenges in NPD, many firms have implemented the Stage-Gate® process to enhance innovation [15, 30]. Among the most critical key success factors in NPD are speed to market, strategy, and tactics [28, 31-33]. Strategy defines the overall direction of innovation efforts, while tactics provide steps to ensure alignment between execution and resource allocation. These two factors are essential for ensuring the innovation process remains focused on speed and efficiency [1, 25, 34]. However, numerous firms continue to grapple with failures that are often attributed to poor organisation and rigorous execution [29, 35, 36].

Cooper developed the Stage-Gate® model in the late 1980s to improve efficiency and effectiveness in product development [3, 37]. The model, which consists of multiple stages separated by gates, initially focused on rigorous planning and evaluation [11].

Despite its success, critics have expressed concern about potential inflexibility in fast-paced environments [9, 38]. One of the primary challenges in systematic innovation is balancing creativity with resource constraints. However, this can be addressed by implementing processes that allow for rapid iteration, granting the organisation greater flexibility [13]. Methodologies such as Agile or Lean can seamlessly integrate into the systematic innovation process, allowing for rapid iterations and continuous hypothesis testing, which improves flexibility in the innovation cycle. By combining the structured approach of Stage-Gate® with the adaptive nature of Agile, firms can better manage uncertainty while maintaining a focus on strategic goals [39]. The Stage-Gate® model's ongoing evolution, including modifications and adaptations, reflects the dynamic nature of the field and the need for continuous refinement to remain competitive [40, 41].

4. The original Stage-Gate® model

Developed on the basis of the 1980’s NewProd system [37, 42], Cooper's original Stage-Gate® model has become a standard framework for managing product development processes. The model consists of distinct stages and gates, each of which serves a distinct purpose in ensuring the project's systematic progression [3].

The five stages of the traditional Stage-Gate® model include scoping, business case development, development, testing and validation, and launch [2]. Each stage consists of high-level tasks and is followed by a gate, which is a decision point at which the project's progress is assessed using predetermined criteria, allowing Go/Kill decisions on further investing [2].

This methodical approach enables innovators and organisations to reduce innovation project failures through planning and control. The model incorporates information gathering, data integration, and analysis at each stage, followed by gates that determine the project's resource investment [2]. Cooper compares this process to purchasing options on an investment, in which low-cost options are acquired initially and subsequent decisions about continuing the investments are made based on increasing levels of information [2].

Each Stage-Gate® stage gathers information to reduce risks and uncertainties [2, 3]. As the project progresses, stages build on one another, requiring different resource allocation at each stage. For example, later stages like testing and validation may require substantial investments in prototypes or market research [14, 15].

The model's reliance on gates ensures that decisions are based on increasingly accurate data, thereby continuously managing risk [43].

The Stage-Gate® model's impact on industry practices has been profound. For instance, large corporations like Procter & Gamble have successfully implemented the model, significantly enhancing their product development efficiency [14, 15]. Other sectors, such as telecommunications, have applied the model to improve portfolio governance [44]. Academic research recognised the model’s value in balancing creativity with control, though some critiques call for industry-specific adaptation [13, 41].

The original Stage-Gate® model begins with the Scoping stage and concludes with the post-Launch assessment [3]. The Discovery Phase, or Idea Stage, is important but not considered a standard stage of the Stage-Gate® model. While the initial stages do not require large financial investments, later stages such as Go to Development involve significant and specific resource allocation [2, 14, 15]. The details of the original Stage-Gate® model are summarised in Table 1.

Table 1 Details of the original Stage-Gate® model

Stage/Gate	Name	Description
Pre-stage, Stage 0	Discovery Phase, Idea Stage	The conception and accumulation of innovative new product ideas.
Gate 1	Idea screen	The selection and priority setting of product ideas for an NPD project regarding a dynamic process with a high degree of uncertainty.
Stage 1	Scoping	A preliminary analysis of the market and technology, including an evaluation of the most fundamental financial values.
Gate 2	2nd screen	A decision regarding the project's progression should be made based on the collection and analysis of information that has been subjected to rigorous conditions.
Stage 2	Building a business case	The conceptualisation of the business case, which includes an in-depth development plan and a launch strategy for the market.
Gate 3	Go to development	The decision must be made regarding the profitability of the project and the release of revered resources.
Stage 3	Development	The development of new technologies as well as the analysis of various marketing and production endeavours.
Gate 4	Go to testing	Evaluation of the project's ability to be technically realised and management of the R&D budget.
Stage 4	Testing and validation	The validation of the financial plan, the evaluation of the technological performances, and the acceptance of the customers all need to be performed.
Gate 5	Go to launch	Authorisation to enter the market.
Stage 5	Launch	Product commercialisation and market entry.
Post-launch review	Monitoring	The launch process is being evaluated.

Source: [3]

However, critics have pointed out potential weaknesses, such as rigidity, which may hinder flexibility in fast-paced industries like high-tech, where rapid iteration and adaptability are crucial [9].

In comparison, Lean or Agile emphasise flexibility and rapid iteration, allowing companies to pivot based on real-time feedback. While the Stage-Gate® model provides structured control, it may not match the speed required in dynamic environments. Nevertheless, it excels in industries where compliance, regulatory considerations, or large-scale investments demand more structured processes, such as deep-tech [16], pharmaceuticals, or consumer goods [2].

In essence, the original Stage-Gate® model is a well-defined and time-tested embedding of clear goals and a competent execution path [2]. In its original form, it has helped improve the efficiency and effectiveness of NPD processes by offering a systematic technique that blends creativity with control [23, 28, 35]. As industries evolve, the need for flexibility has led to modern adaptations [13, 26, 39].

5. Evolution of the original Stage-Gate® model

Introduced by Cooper in 1990, the Stage-Gate® model [3] emerged as a distinct five-stage, five-gate framework, building on the insights and methodologies of the earlier NewProd system [37, 42]. The original version adopted a 'one size fits all' approach for a structured and systematic innovation management [3].

While the NewProd system laid the groundwork, the introduction of the Stage-Gate® model marked a significant advancement in the field, providing a clear and standardised process that would become widely adopted across various industries [1, 2, 40, 45]. However, users of the original Stage-Gate® model found its rigid five-stage structure limiting for smaller projects, leading to early adaptations and community customisations that recognised the need for flexibility and responsiveness in varying project contexts [1].

Table 2 The Stage-Gate® model evolution showcasing the key developments and adaptations

Year	Name	Short description
1985	NewProd	NewProd, an industrial new product development process model with seven stages (Idea, Preliminary Assessment, Concept, Development, Launch, Trial, and Launch), includes activities and evaluation points for product development and marketing, emphasising market orientation and timely evaluation [37, 42].
1990	Stage-Gate process	Stage-Gate® enhances efficiency in product development from idea to launch, treating innovation as a manageable process with stages and gates as quality checkpoints. This improves decision-making, focus, and speed [3].
2008	Stage-Gate LITE and XPRESS (Spiral Development)	Spiral development allows quick iteration and design adjustments. LITE and XPRESS versions scale the process for different project types; LITE handles simple requests, and XPRESS addresses moderate-risk projects like improvements, modifications, and extensions [2, 45].
2016	Agile-Stage-Gate	The Agile-Stage-Gate model combines Stage-Gate® structure with Agile methodologies to enhance response times, communication, and productivity, helping industries launch products faster [40, 46].
2022	5th generation Stage-Gate (Triple A System, Value Stream Management)	The 5th Generation Stage-Gate® Idea-to-Launch Process is comprehensive and a flexible NPD system that enhances efficiency, effectiveness, and success with additions like Value Stream Mapping, Concurrent Processing, Iterations, Tougher Gates, and Agile integration are added to improve NPD [1].

5.1 Hybridisation of the Stage-Gate® model

Much like species in nature, the Stage-Gate® model has evolved to adapt in complex innovation processes. From its origins as a standard, one-size-fits-all framework [3], it initially underwent small incremental adjustments designed to address the specific needs of different projects [2].

As industries became more complex and innovation processes more dynamic, the model experienced more significant adaptations. New branches of the original Stage-Gate® system emerged–frameworks that shared the same fundamental principles but introduced new, hybrid approaches to development [40, 46]. These new hybrids represent major evolutionary leaps, transforming how companies approach product development. By incorporating iterative feedback, rapid prototyping, and a customer-centric focus, these hybrids have significantly expanded the model's adaptability to a broader range of industries, particularly those requiring continuous market responsiveness and flexibility [1, 13].

As Table 2 illustrates, the Stage-Gate® model has transformed over the years to meet evolving industry demands and challenges with modifications addressing various challenges [1, 2, 5, 45]. Some organisations have adapted the Stage-Gate® model to suit specific needs, either by reducing the number of stages to speed up simpler projects [2, 16] or by increasing them for more complex ones, ensuring stringent quality controls [1, 47]. Adaptations have also focused on accelerating product development through parallel processing [46], spiral development [2], and integrating continuous customer feedback [1]. Sector-specific customisations have been introduced to keep up with rapid technological advances and meet ever faster market demands [1, 12, 40, 47].

Most adaptations redefine stages or gates, while some expanded the model by introducing entirely new stages that combine traditional Stage-Gate® principles with other innovation methodologies [5, 8, 13, 16]. Agile principles marked a major turning point [2, 40], leading to hybrid models that integrate also other innovation methodologies such as Design Thinking and Lean Startup [12, 13].

This multifaceted approach has enriched the Stage-Gate® model, providing managers with insights into the combinatory possibilities of different methodologies [1, 28, 45]. It enables informed decisions, accelerating NPD [28] and represents a significant milestone in the evolution of the Stage-Gate® model, reflecting a growing recognition of the need for flexibility, customisation, and responsiveness in innovation management [1, 28, 45].

In today's dynamic innovation landscape, flexibility and adaptation remain essential [48, 49]. While the Stage-Gate® model has proven robust and widely adopted, it is not immune to the unique challenges presented by various industries and projects [16]. As a result, the model's stages and gates have undergone conscious adaptations – referred to as customisation or hybridization – to meet diverse industry needs [2, 13, 40].

Customisation extends beyond the integration of methodologies like Agile, Design Thinking, and Lean Startup; reflecting broader shifts in NPD and innovation demands [12]. In industries such as manufacturing and healthcare, process customisation is essential for coping with rapid technological advances and meeting regulatory standards [1, 50, 51]. For instance, the healthcare industry has integrated additional gates for regulatory approvals, ensuring compliance with strict medical standards before advancing to development stages [16, 50].

However, customisation brings both benefits and challenges. While customisations enhances the responsiveness and adaptability of the Stage-Gate® model, it can also introduce complexities that must be carefully managed [14, 15, 39]. One of the primary risks of over-customisation is added complexity, which can slow down decision-making and reduce the model's original efficiency customisation [7, 13]. In the technology sector, this has meant combining stages to speed up development [52], while in healthcare, additional gates may ensure regulatory compliance [47, 50]. Balancing flexibility, customisation, and the core principles of the Stage-Gate® model can be challenging [39, 53].

5.2 Stage-Gate® model hybrids

Building on the evolution of the Stage-Gate® model, several hybrid approaches have emerged that integrate other innovation methodologies to enhance flexibility, responsiveness, and customer-centricity. These hybrids represent significant adaptations of the original model, combining its structured framework with the iterative and collaborative principles of methodologies like Agile, Design Thinking, Lean Startup, and TRIZ.

One of the most well-known hybrids is the integration of Agile principles into the Stage-Gate® model [1]. Agile methodology divides the development cycle into sprints, which are characterised by iterative development, continuous feedback, and adaptability [36]. This hybrid approach combines the flexibility and rapid iteration of Agile with the structure and discipline of the Stage-Gate® model. Cooper & Sommer [40, 46] report positive results from this integration.

In software and technology industries, the Agile-Stage-Gate® model improves time-to-market by enabling faster product iterations without sacrificing strategic oversight [9]. Teams can rapidly prototype and test concepts, allowing for quick development cycles and constant product refinement. The Agile-Stage-Gate® model enhances collaboration by encouraging cross-functional teamwork and breaking down departmental silos that often hinder innovation. It also improves responsiveness to market changes, enabling swift adaptations to shifts in customer demands. Communication becomes more streamlined, with clear feedback loops enhancing understanding among team members and stakeholders.

However, integrating Agile with Stage-Gate® comes with challenges. Aligning Agile's iterative cycles with Stage-Gate®'s sequential decision-making process can cause tension between teams or team members focused on flexibility and those adhering to predefined milestones. Overcoming this requires careful planning and open communication to ensure both approaches complement each other. Success stories from industries such as consumer electronics highlight significant reductions in development time when these challenges are effectively managed [12, 16].

Design Thinking, a human-centred innovation methodology, emphasises empathy, creativity, and iterative problem-solving [54]. When integrated with the Stage-Gate® model, it fosters a user-centric approach to product development, blending innovation with structured development processes [53]. Teams focus on understanding and addressing user needs, which encourages creative ideation and exploration of a wider range of solutions.

Design Thinking unfolds through five stages: Empathize, Define, Ideate, Prototype, and Test. This integration enables rapid experimentation, prototyping, and iterative testing for efficient idea validation. It also promotes interdisciplinary collaboration, bringing together diverse perspectives to enrich the development process. Balancing creativity with systematic development ensures that products align with real user demands, guided by continuous feedback [55].

Maintaining the creative freedom that Design Thinking encourages while adhering to the structured, sequential gates of the Stage-Gate® model, poses certain challenges [12]. Organisations need to foster an environment where innovation thrives within a disciplined framework. This approach has been particularly successful in industries like consumer goods, where deep understanding of user preferences is essential [16, 56].

Integrating Lean Startup principles introduces a business-focused, iterative, customer-centric, and experimental approach [12, 57]. This hybrid emphasises market alignment, efficiency, and reduced resource waste [58] by focusing on building minimum viable products (MVPs), conducting rapid testing, and learning from customer feedback to pivot accordingly [16, 58-60]. By minimising waste and optimising resource utilisation, organisations achieve efficiency and prioritise real customer needs throughout the development process. Rapid iterations enable continuous learning and swift incorporation of real-world insights, ensuring the product remains aligned with market needs [58, 59, 61].

While highly effective in industries like software development, challenges arise when applying Lean Startup principles in heavily regulated sectors like pharmaceuticals, where compliance demands a more structured approach [53, 57, 59, 62].

The Theory of Inventive Problem Solving (TRIZ), developed by Genrich Altshuller, is a methodology that offers a systematic approach to solving engineering and design challenges, offering tools for creative problem solving, overcoming contradictions, and inventing new solutions [63]. Integrating TRIZ with the Stage-Gate® model combines systematic problem-solving with structured development. This encourages teams to challenge assumptions, think creatively, and develop breakthrough innovations within a methodical and disciplined framework [64].

While the hybrid promotes inventive problem-solving and robust decision-making, balancing TRIZ's analytical rigour with early-stage flexibility can be challenging. Teams must navigate the depth of analysis required by TRIZ without hindering the speed and adaptability necessary in fast-moving sectors [10, 63, 65-67]. A summary of the key benefits of each Stage-Gate® model hybrid is provided in Table 3.

Table 3 Summary of the key benefits of each Stage-Gate® model hybrids

Hybrid Variant	Key Benefits
Agile-Stage-Gate® model [1, 9, 12, 16, 36, 38-40, 46]	– Rapid prototyping and development – Continuous improvement of the product – Enhanced cross-functional collaboration – Quick responsiveness to market changes – Streamlined communication and feedback loops – Improved risk management – Alignment with customer needs
Design Thinking and the Stage-Gate® model [12, 16, 53, 56]	– User-centric innovation through empathy – Enhanced creativity and ideation – Rapid experimentation with prototypes – Interdisciplinary collaboration – Balanced structure between creativity and systematics – Increased market alignment – Feedback-oriented development
Lean Startup Approach and the Stage-Gate® model [12, 57-59, 61]	– Cost efficiency by minimising waste – Strong customer alignment – Accelerated learning through rapid iterations – Market responsiveness to trends – Risk mitigation via continuous validation – Enhanced scalability – Data-driven decision-making based on real-world insights
TRIZ and the Stage-Gate® model [24, 63-65, 67-70]	– Encourages inventive problem-solving – Provides a systematic approach to complex challenges – Facilitates the generation of novel ideas – Help resolve conflicting requirements – Integrates knowledge across industries – Aligns innovations with business goals – Supports robust, analytical decision-making

5.3 Implications of Stage-Gate® model hybridisation

Customisation has proven critical in maintaining the relevance and effectiveness of the Stage-Gate® model in an ever-changing landscape [1]. These hybrid approaches – Agile, Lean, Design Thinking, and TRIZ – offer distinct advantages but also introduce complexities that require careful management, especially when resources are limited [13, 16, 39, 62].

As industries evolve, balancing flexibility with structured decision-making is essential. Customisation frameworks help organisations identify key decision points, assess risk levels, and tailor the model to align innovations with strategic goals, avoiding over-complication [1, 13, 53, 59].

The evolution of the Stage-Gate® model through hybrids reflects the recognition that no one-size-fits-all solution exists [23, 28, 35]. Leveraging each methodology's strengths, organisations can create a more adaptive, customer- or industry-centric, and responsive innovation process, enabling them to stay competitive in today's rapidly changing markets [8, 12, 26].

This need for adaptability is particularly evident in the early stages of innovation, when ideas take shape and market alignment begins. Focusing on adaptability during these initial phases allows organisations to generate numerous ideas, explore a wider range of concepts, respond swiftly to emerging trends, and integrate customer feedback more effectively. Emphasis the front end of innovation (FEI) helps set a solid foundation for the product development process, increasing success in later stages [2, 13, 25].

6. The neglected stage

The role and importance of the front end of innovation (FEI), have gained increasing attention in the last decade [4, 5, 13, 71]. This NPD stage involves the generation of ideas and their preliminary analysis, forming the foundation of subsequent product development phases [72].

Research consistently emphasises that a well-managed FEI can significantly impact the success of new products. Florén and Frishammar [4] highlighted the FEI's role in capturing customer insights and turning them into innovative product concepts, leading to sustainable competitive advantage. Markham [71] noted that neglecting the FEI often leads to cost overruns and delays, as initial uncertainties and risks remain unaddressed. Cooper [1] emphasised that robust FEI processes lead to higher NPD success rates through better market alignment and resource allocation.

Building on this foundation, recent research continues to validate and expand upon these findings. Kock *et al*. [73] demonstrated that effective ideation portfolio management in the FEI enhances innovation performance through better idea selection and prioritization. Eling *et al*. [74] found that combining rational and intuitive approaches in early idea evaluation improves decision quality and NPD outcomes. Koen *et al*. [75] confirmed that organisations with effective FEI processes achieve faster time-to-market and greater product success. Moreira and Vidor's [13] bibliometric analysis showed that emphasising FEI shortens development cycles and improves product performance.

Despite this consensus, the Stage-Gate® model underemphasises the critical role of the FEI. While, Moreira & Vidor highlight that while the FEI is essential, the original model treats it as a preliminary stage, referring to it as “Idea Stage “, "Stage 0" or the "pre-work" phase [76]. This framing diminishes the importance of Discovery Phase in practice, even though it directly influences the success of product innovation by shaping initial concepts and aligning them with organisational goals [13, 75].

This paper proposes a redefinition of the Stage-Gate® model, renumbering Stage 0, known as Discovery Phase, to Stage 1 and shifting all subsequent stages accordingly. This change emphasises the importance of the initial idea generation phase and aligns with the views expressed in existing research. By renumbering this phase as "Stage 1", we acknowledge that it is not a mere prelude but a critical phase that sets the tone for the entire innovation process [1, 4, 5, 71].

Renumbering the Discovery Phase addresses the inconsistent treatment of the FEI across industries, despite evidence of its pivotal role in determining the outcome of NPD projects [13]. Renaming the Discovery Phase or Idea Stage as Stage 1 would align the Stage-Gate® model with this evidence, placing greater emphasis on idea exploration and early-stage evaluation. While Cooper [2, 76] describes the fuzzy front end (including ideation, scoping, and building the business case) in his work, he did not explicitly reclassify it as Stage 1. This research proposal offers an original contribution by advocating for its formal recognition as a critical component rather than a preliminary phase.

This shift also addresses emerging challenges in NPD, such as sustainability and data-driven decision-making, which are becoming increasingly important [13, 26]. The FEI should incorporate a more formalised evaluation of environmental and social impacts, as well as comprehensive data analysis [14, 15]. Though not fully addressed in the current Stage-Gate® model, integrating these considerations into Stage 1 ensures that innovations are aligned with modern expectations and societal goals at inceptions [7, 13, 16].

By treating the FEI as Stage 1, the Stage-Gate® model becomes even more adaptable, allowing for the seamless integration of contemporary innovation methodologies. The redefined model fosters flexibility and ensures that organisations can tailor the model to suit their specific project needs and industry trends. The following innovation methodologies can be integrated at different stages of the revised Stage-Gate® model:

- *Design Thinking:* Integrating Design Thinking into Stage 1 encourages creativity and empathy-driven problem-solving from the very beginning of the innovation process. This methodology focuses on understanding user needs, generating ideas, and prototyping solutions—activities that are essential during the Discovery or Idea phase. By embedding Design Thinking at this early stage, organisations can ensure that user-centric insights drive the development of ideas, setting a strong foundation for all subsequent stages [12, 16, 55, 56].
- *TRIZ:* Traditionally applied in later stages, TRIZ can also benefit Stage 1 by fostering systematic innovation from the outset. TRIZ provides a structured approach to solving complex design and engineering challenges, making it an effective tool for overcoming contradictions and identifying inventive solutions early in the ideation process. Applying TRIZ in

Stage 1 helps gather valuable insights and shape ideas that are not only creative but also technically viable, streamlining the path to later-stage development [64, 65, 67].

- *Agile:* Commonly utilised in Stages 4 and 5 for iterative development, Agile practices enhance Stage 1 by introducing iterative cycles of ideation and feedback right from the start. In the Discovery or Idea Stage, Agile's flexibility and focus on continuous feedback ensure that ideas evolve rapidly in response to market needs or stakeholder input. This early incorporation of Agile principles helps reduce risk by refining concepts before they move into more resource-intensive stages [38, 39, 46].
- *Lean Startup:* Though usually applied in Stages 3 and 6, Lean Startup can also be integrated into Stage 1 to emphasise rapid experimentation and customer validation from the earliest phases of innovation. By testing hypotheses and gathering feedback early on, organisations can ensure that the ideas they pursue are aligned with market demands, thus reducing the likelihood of costly pivots in later stages. Lean's focus on minimising waste and continuous validation makes it particularly effective in ensuring that the ideas generated during Stage 1 are viable and scalable [12, 57-59, 61].

Incorporating these innovation methodologies within the appropriate stages strengthens the Stage-Gate® model, making it more comprehensive and adaptable. This approach aligns the innovation process with modern industry demands, ensuring that each stage – starting with the newly emphasised Stage 1 – captures the strategic and operational needs of organisations more effectively. As the model adapts to different project types and sectors, it becomes a more powerful tool for managing complexity and fostering continuous innovation. The renumbered Stage-Gate® model, with its increased focus on the FEI, better equips organisations to navigate competitive and fast-evolving markets. This is evolutionary step towards sustaining innovation success and maintaining a competitive edge [1, 13, 16].

6.1 Representation of the renumbered Stage-Gate® model

To clarify the proposed renumbering and restructuring, the Table 4 summarises the changes. Renumbering as Stage 1, and shifting all subsequent stages accordingly, provides clearer guidance and emphasises FEI's integral role.

Table 4 Details of the amended Stage-Gate® model

Stage/Gate	Name	Description
Stage 1	Idea Stage	Conception and accumulation of innovative new product ideas.
Gate 1	Idea screen	Selection and prioritisation of product ideas for an NPD project within a dynamic, uncertain process.
Stage 2	Scoping	Preliminary analysis of the market and technology, including basic financial evaluation.
Gate 2	2nd screen	Decision on project progression based on rigorously analysed information.
Stage 3	Build a business case	Conceptualisation of the business case, including an in-depth development plan and a launch strategy for the market.
Gate 3	Go to development	Decision on project profitability and resource allocation.
Stage 4	Development	Development of new technologies as well as the analysis of various marketing and production endeavours.
Gate 4	Go to testing	Evaluation of technically feasibility and R&D budget management.
Stage 5	Testing and validation	Validation of the financial plan, evaluation of the technological performance, and customer acceptance.
Gate 5	Go to launch	Market entry authorisation.
Stage 6	Launch	Product commercialisation and market entry.
Post-launch review	Monitoring	Evaluation of the launch process.

The Discovery Phase or Idea Stage serves a critical role in the innovation process. To further substantiate the argument for renumbering this stage as Stage 1, we look to real-world examples where this stage has already been implicitly or explicitly recognised as crucial in various industries. The following case studies provide evidence of how leading organisations have adapted the Stage-Gate® model, supporting the proposed renumbering and its relevance in fostering successful innovation processes.

7. Case studies

Recognising the Discovery Phase as Stage 1 in the Stage-Gate® model is not just theoretical. Real-world applications across industries have implicitly or explicitly demonstrated this stage's crucial role in innovation. This section presents case studies that show how different organisations have adapted the model, supporting the necessity and practical impact of renumbering.

7.1 Case study 1: The Fiat Mio crowdsourcing project in Brazil

The Fiat Mio crowdsourcing project presents a compelling case for redefining the Discovery Phase and Idea Stage as Stage 1. The project began when a Fiat executive recognised the need to better respond to consumer demands. This led to the development of a co-creation platform that engaged the public in conceptualising a new car [17].

Fiat started with "Original Idea" stage, setting the foundation for the project's innovation process. Redefining the Discovery phase as "Stage 1" in the Stage-Gate® model aligns with Fiat's approach, emphasising the importance of this early ideation stage in fostering innovation [17]. As the project progressed, the remaining five stages focused on development, testing, and commercialisation, following a modified version of the Stage-Gate® model.

Fiat's approach shows how integrating consumer input early in the process provided insights that shaped the product from the start. Treating idea generation as a formal phase, Fiat used structured crowdsourcing to refine ideas in real time, demonstrating early-stage collaboration's value. Additionally, it reflects the growing importance of consumer-centric innovation practices in modern industries.

The Fiat Mio project validates redefining the Discovery Phase, illustrating how early-stage ideation, supported by crowdsourcing, is critical to NPD [17].

7.2 Case study 2: Agile-Stage-Gate® hybrid in the toys and power segments

Sommer *et al.* [38] presents a comparative case study of two companies in the toys and power segments that adopted hybrid Agile-Stage-Gate® models, using "Idea" stage as Stage 1, explicitly recognising it as the initial stage. Recognising "Idea" stage as Stage 1 highlighted its importance for aligning development with market needs.

Combining Agile with the Stage-Gate® model enhanced flexibility and responsiveness as despite challenges these companies resulted in greater adaptability and faster decision-making in the early stages of innovation [38]. This approach shows that treating the early idea phase as a formal stage can reduce the time and cost of later stages by catching potential problems early on. By implementing Agile cycles of feedback, companies were able to ensure that Stage 1 was not merely a brainstorming session but a rigorously managed phase critical to the success of the entire product development process.

Both companies used iterative feedback mechanisms within Stage 1 to adapt to changing market conditions [38]. This operationalisation of the Discovery Stage reflects the need to redefine it as Stage 1, showing that modern NPD depend on the flexibility and responsiveness that Agile methodologies can bring to the early stages of innovation.

7.3 Case study 3: Procter & Gamble's SIMPL model

Procter & Gamble (P&G), a leader in product innovation, exemplifies the prioritisation of early-stage innovation through its adaptation of the Stage-Gate® model, called SIMPL [52]. The SIMPL consists of five stages and four gates, with Stage 1 being explicitly named Discovery:

- Stage 1: Discovery – This stage focuses on ideation, exploration, and the identification of new opportunities, setting the foundation for subsequent stages.

- Stage 2: Design – Conceptualising and designing products to align with customer needs and market demand.
- Stage 3: Qualify – Rigorous testing and validation to ensure product quality.
- Stage 4: Ready – Preparing the product for launch, including supply chain and marketing coordination.
- Stage 5: Launch – Product release, including sales and distribution efforts.

P&G's formal implementation of the Discovery Stage as Stage 1 underscores the importance of early-stage innovation. P&G's framework rigorously evaluates new opportunities to ensure alignment with market needs and strategic goals before progressing, enhancing the likelihood of product success. This operational approach further validates the necessity of redefining the Discovery Phase in the broader Stage-Gate® model [52].

Moreover, P&G's SIMPL model provides a clear example of how Stage 1 can be tailored to meet specific organisational needs, depending on the type of product or market being targeted. This structured yet flexible approach demonstrates that redefining the Discovery Phase is more than symbolic; it is essential for effective innovation management. This case study supports the argument that emphasising Stage 1 helps align ideation with business objectives.

7.4 Case study 4: Sustainability in the I2P^3® process at Evonik Industries

Wojciechowski *et al*. [77] describe how Evonik Industries AG implemented the I2P^3® Process, a sustainability-focused innovation framework based on the Stage-Gate® model. It includes six stages, starting with Stage 1: Idea Development, which assesses the societal and environmental impact of new ideas through the dimensions of People, Planet, and Profit.

Evonik's approach ensures sustainability is embedded from the start, evaluating ideas not just for ideation but for their impact on people, the environment, and profitability and solidifies the argument that the Discovery Phase or Idea stage should be redefined as Stage 1 to reflect its integral role in assessing sustainability, aligning with the Stage-Gate® model's goals of balancing innovation with societal and environmental responsibility [77].

Evonik's I2P^3® Process demonstrates how early-stage innovation can incorporate sustainability considerations from the very beginning. The company uses Stage 1 not only for ideation but also as a phase for conducting comprehensive sustainability assessments, evaluating each new idea based on its potential impact on people, the planet, and profitability [77]. This approach ensures that sustainability is not an afterthought but an integral part of the product development process.

By formalising this phase, Evonik rigorously scrutinises ideas before advancing them, aligning innovations with market and corporate social responsibility goals. This case supports treating the Discovery Stage as a formal NPD step to meet the growing demand for sustainable, responsible innovation.

7.5 Contextual evidence of discovery phase variation across sectors

The Discovery Phase is not a one-size-fits-all template. Companies vary its depth, tools, and resources according to five context drivers:

- *Resource & knowledge intensity:* industries vary in how much they invest in R&D, particularly during early-stage innovation. High-index sectors, such as electronics and pharmaceuticals, typically allocate more resources to basic, original and high-quality innovation efforts [78]. Cross-national evidence confirms that greater R&D investment, particularly in early phases, is positively associated with enhanced innovation performance across sectors and economies [34, 78].
- *Regulation & compliance:* sectors with high regulatory demands, such as pharmaceuticals, embed many more checkpoints early in the Discovery Phase to ensure GMP alignment and technical feasibility for later EMA/IND filings [79]. This increased diligence extends early-stage activities: preclinical development durations grew by 17 % between 2004 and 2012 [27]. This early diligence improves readiness for downstream development stages by strengthening de-

cision making and reducing the risk of late stage project termination, an increasingly critical need given that up to 90 % of drug candidates still fail in clinical trials [80].

- *Methods applied in early-stage innovation:* the Discovery Phase acts as a performance lever across sectors, particularly when supported by systematic methods and aligned with downstream development stages, reinforcing its role as a strategic driver of innovation success [50]. In software and IT, Agile practices such as iterative sprints and early customer feedback are integrated early, improving business case clarity and reducing risk. This approach is linked to gains in speed, cost, and quality [36], as well as a 25 % reduction in project effort and 20 % less rework in hybrid Agile Stage Gate models [46] In discrete manufacturing, simulation-based planning is increasingly embedded into the Discovery Phase to support early feasibility decisions. A print-shop pilot increased weekly output by 52 % and cut job time in half by using pre-execution modelling to optimise production scenarios [81].
- *Intellectual-property risk:* high-IP sectors such as deep-tech and specialty chemicals embed patent landscaping, Freedom-to-Operate and other IPR related checks into the Discovery Phase. These early evaluations begin with idea screening and intensify through early gates, adding time and coordination effort, but improve decision quality significantly and reduce infringement risk in later stages [21, 78, 82].
- *AI integration:* Data-rich industries are increasingly embedding generative AI into the Discovery Phase to support ideation, customer research, and feasibility assessment. Electronics and chemical companies use these tools to auto-rank ideas, simulate synthesis routes, and reduce uncertainty. Industry pilots report a 25-35 % drop-in screening time and more consistent Go/Kill decisions, aligning with broader digital economy trends accelerating innovation in manufacturing [83, 84].

These examples confirm that, although companies retain the same gate logic, they adjust the Discovery Phase to align with their resource intensity, regulatory requirements, preferred methods, intellectual property risk, and level of AI readiness. Taken together, this evidence reinforces our argument that the Discovery Phase, previously referred to as Stage 0 or the Idea Stage, is a crucial part of the innovation process and should be renumbered as Stage One to reflect its central role in driving successful innovation.

8. Discussion

This paper has proposed a significant refinement of the Stage-Gate® model by renumbering the Discovery Phase as Stage 1, recognising its essential role in the innovation process [1]. This change is not a superficial adjustment but rather a strategic realignment that simplifies the model and underscores the significance the importance of ideation and creativity [5, 13] within the Stage-Gate® model.

By renumbering the Discovery Phase, the model becomes more comprehensive and better aligned with modern innovation methodologies such as Agile, Design Thinking, TRIZ, and Lean Startup [1, 16, 26]. Case studies like the Fiat Mio crowdsourcing project and SIMPL by Procter & Gamble demonstrate that these adjustments improve clarity, flexibility, and strategic focus in the NPD process [17, 52].

The paper shows that ideation has always been a crucial component of NPD, and findings in FEI research confirm that we should never view it as optional or peripheral, but rather as the cornerstone of successful innovation. Modern methodologies further highlight the centrality of the Discovery Phase by integrating it into flexible and iterative innovation frameworks. For instance, Agile and Lean Startup approaches stress continuous feedback and adaptability, which aligns with the early-stage ideation process critical to innovation [1, 13, 16, 26].

While the benefits are clear, it is important to acknowledge the challenges that come with this change. Emphasising the Discovery Phase as Stage 1 could risk overloading this phase with too many tasks and expectations, which might slow down the early stages of the NPD process. The increased complexity of managing ideation, opportunity exploration, and market analysis as formalised steps could result in longer timelines, especially for companies that need to iterate quickly [13]. Additionally, there is the risk that, by formalising the Discovery Phase, some companies might become too rigid in their early-stage exploration, which could stifle creativity and hinder the flexibility that innovation requires [7].

The shift to renumber the Discovery Phase as Stage 1 provides a more intuitive, holistic, and consistent framework for innovation management, emphasising the critical role of early-stage ideation in driving the development of new products. However, this formalisation also introduces a potential challenge: balancing the need for structured, organised, orchestrated actions with the creative freedom required for ideation. Over-structuring the Discovery Phase may limit the open-ended exploration that often leads to breakthrough innovations [1].

Emerging trends like artificial intelligence (AI), big data analytics, and sustainability practices will likely shape the future evolution of the Stage-Gate® model [1]. These trends offer significant opportunities to enhance the model's flexibility, efficiency, and innovation capabilities, but they also present challenges that organisations must address.

AI and big data, for example, can revolutionise early-stage decision-making by providing predictive insights into market trends and customer preferences, transforming the Discovery Phase [5, 13]. Similarly, sustainability practices may lead to the inclusion of new gate parameters that focus on evaluating the environmental and societal impacts [16, 26, 85]. Meanwhile, the continued rise of iterative innovation methodologies will drive the need for rapid market responses, particularly in industries with fast-moving dynamics [1, 40, 46].

However, integrating these emerging trends presents both challenges and opportunities. One challenge is ensuring that the adoption of these improvements does not over-complicate the model, which could reduce its overall efficiency or introduce decision-making delays [7, 13]. On the other hand, the ability to adapt more seamlessly to technological and regulatory changes will offer a significant competitive advantage [1].

Small and medium enterprises (SMEs) in particular may face different challenges when implementing a systematic innovation process compared to larger corporations [86]. Limited resources, budget constraints, a lack of innovation competencies, and less formalised processes can impact their ability to follow a structured NPD approach like Stage-Gate® [1]. However, these firms can benefit from a more flexible, iterative model that still emphasises early-stage validation and alignment with business goals [13].

The adoption of a revised Stage-Gate® model, starting with the Discovery Phase, enables organisations to rethink their innovation processes. This change, along with modern innovation methodologies, can boost the success rates of NPD in fast-moving industries where speed to market is crucial. Integrating data-driven insights from AI into Stage 1 will not only improve idea validation but also enhance information flow, risk management, and decision-making throughout all stages of the innovation process [13]. As industries and technologies continue to evolve, the revised Stage-Gate® model will remain adaptable, offering a flexible yet structured approach to innovation. Moving forward, the integration of emerging trends such as AI, big data, and sustainability will further enhance the model’s relevance and applicability, ensuring that it remains a vital tool for managing innovation in an ever-changing business landscape.

9. Conclusion

The continuous evolution of the Stage-Gate® model reflects the growing need for customisation and flexibility in response to emerging trends and industry demands. This paper has proposed a fundamental shift in the model’s structure by renumbering the Discovery Phase as Stage 1, recognising its essential role in the innovation process. This change aligns the model with modern innovation methodologies and emphasises the importance of early-stage ideation and creativity.

For practitioners, this renumbering enhances clarity and focus during the early stages of product development. It facilitates cross-functional collaboration, decision-making, and alignment with methodologies like Agile and Design Thinking, fostering a more flexible and customer-centred innovation process. Policymakers should support the adoption of such customisations, especially in industries where rapid innovation cycles and early-stage ideation are critical to market success.

While this paper makes a strong argument for renumbering the Discovery Phase, we must acknowledge several limitations. Firstly, empirical validation is required to assess the impact of this change across a broader spectrum of industries and company sizes. Further research is necessary to explore the unique challenges that different sectors may encounter when adapting the revised model. Additionally, it is important to recognise that for some organisations, this adaptation may not be entirely novel. Certain companies may already be informally integrating Discovery Phase activities into other stages of the model. These companies may compensate for not having a distinct Discovery Phase by using innovation techniques and methodologies within other stages, thus performing early ideation and exploration implicitly. Future research should test whether companies are indeed engaging in such implicit activities and assess whether formalising the Discovery Phase would offer them additional benefits.

Secondly, the increased complexity introduced by this customisation must be carefully managed to ensure that the Stage-Gate® model retains its core efficiency. Over-complicating the model could lead to decision-making delays, slowing down innovation rather than enhancing it. The balance between structured processes and maintaining flexibility for creative exploration will also be crucial in determining the model's success.

The proposed changes and adaptations of the Stage-Gate® model open new avenues for empirical research and practical exploration. Future research should focus on:

- *Implicit innovation activities:* research should examine how companies that do not formally recognise the Discovery Phase still engage in early ideation and exploration implicitly through other stages of the model. It is crucial to test whether such companies use innovation techniques and methodologies in these stages, which effectively substitute for a distinct Discovery Phase. Validating these practices will help determine whether formalising the Discovery Phase offers tangible benefits over implicit approaches [13].
- *Integration with other models:* exploring how the Stage-Gate® model can be effectively integrated with other innovation and project management models, such as TRIZ, Design Thinking, or Lean Startup, to create more holistic and flexible frameworks [1, 9, 13, 38, 40].
- *Role of AI in decision-making*: investigating the role of AI in enhancing decision-making, particularly in the early stages of innovation. AI could provide more data-driven insights into market trends, customer preferences, and project feasibility [1, 13, 16].
- *Impact Innovation:* examining how sustainability practices can be more deeply embedded within the Stage-Gate® model. Research could concentrate on creating new gate parameters, gates, and stages that evaluate environmental and societal impacts, prioritising sustainability throughout the innovation process [13, 26, 39].
- *Cross-industry applicability:* Investigating the impact of these proposed changes across different industries, considering how sector-specific challenges may affect the integration of hybrid models and customisation efforts [46, 49, 59].
- *Methodological approaches:* Future research on the Stage-Gate® model's evolution should employ a variety of research methods, including case studies, experimental design, surveys, interviews, risk and resource considerations [87]. Cross-industry analysis would provide valuable insights into how different sectors customise the model to meet their specific needs, offering a more comprehensive understanding of the model's adaptability and flexibility.

The flexibility and adaptability of the Stage-Gate® model have long been central to its widespread adoption across industries. By renumbering the Discovery Phase and incorporating emerging technologies, trends, and methodologies, this revised model holds the potential to significantly enhance the success of NPD processes. Given the increasing complexity of innovation processes, it is essential that further empirical research is conducted to validate these changes and assess their broader implications. The model's ability to continuously evolve in alignment alongside new innovation practices will ensure that it remains a powerful and relevant tool for organisations navigating the complexities of modern NPD.

References

[1] Cooper, R.G. (2022). The 5-th generation Stage-Gate idea-to-launch process, *IEEE Engineering Management Review*, Vol. 50, No. 4, 43-55, doi: 10.1109/emr.2022.3222937.

[2] Cooper, R.G. (2008). Perspective: The Stage-Gate® idea-to-launch process–update, what's new, and NexGen systems, *Journal of Product Innovation Management*, Vol. 25, No. 3, 213-232, doi: 10.1111/j.1540-5885.2008.00296.x.

[3] Cooper, R.G. (1990). Stage-gate systems: A new tool for managing new products, *Business Horizons*, Vol. 33, No. 3, 44-54, doi: 10.1016/0007-6813(90)90040-I.

[4] Florén, H., Frishammar, J. (2012). From preliminary ideas to corroborated product definitions: Managing the front end of new product development, *California Management Review*, Vol. 54, No. 4, 20-43, doi: 10.1525/cmr.2012.54.4.20.

[5] Bhatia, A., Cheng, J., Salek, S., Chokshi, V., Jetter, A. (2017). Improving the effectiveness of fuzzy front end management: Expanding Stage-Gate methodologies through agile, In: *Proceedings of 2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, USA, 1-8, doi: 10.23919/picmet.2017.8125390.

[6] Masters, B., Thiel, P. (2014). *Zero to one: Notes on start ups, or how to build the future*, Random House, New York, USA.

[7] Bansal, P.T., Grewatsch, S. (2020). The unsustainable truth about the Stage-Gate new product innovation process, *Innovation*, Vol. 22, No. 3, 217-227, doi: 10.1080/14479338.2019.1684205.

[8] Smolnik, T., Bergmann, T. (2020). Structuring and managing the new product development process-review on the evolution of the Stage-Gate® process, *Journal of Business Chemistry*, Vol. 17, No. 2, 41-57, doi: 10.17879/22139478907.

[9] Cooper, R.G., Sommer, A.F. (2020). New-product portfolio management with agile, *Research-Technology Management*, Vol. 63, No. 1, 29-38, doi: 10.1080/08956308.2020.1686291.

[10] Liu, X.W., Wang, X.Z., Lyu, L.C., Wang, Y.P. (2022). Identifying disruptive technologies by integrating multi-source data, *Scientometrics*, Vol. 127, No. 9, 5325-5351, doi: 10.1007/s11192-022-04283-z.

[11] Cooper, R.G. (2011). Perspective: The innovation dilemma: How to innovate when the market is mature, *Journal of Product Innovation Management*, Vol. 28, No. s1, 2-27, doi: 10.1111/j.1540-5885.2011.00858.x.

[12] Cocchi, N., Dosi, C., Vignoli, M. (2021). The hybrid model matrix enhancing Stage-Gate with design thinking, lean startup, and agile, *Research-Technology Management*, Vol. 64, No. 5, 18-30, doi: 10.1080/08956308.2021.1942645.

[13] Moreira, L.F., Vidor, G. (2024). The importance of the product development process for companies: A bibliometric analysis of the Stage-Gate model, *Paradigm*, Vol. 28, No. 1, 7-25, doi: 10.1177/09718907241243188.

[14] Cooper, R.G., Edgett, S.J., Kleinschmidt, E.J. (2002). Optimizing the Stage-Gate process: What best-practice companies do—II, *Research-Technology Management*, Vol. 45, No. 6, 43-49, doi: 10.1080/08956308.2002.11671532.

[15] Cooper, R.G., Edgett, S.J., Kleinschmidt, E.J. (2002). Optimizing the Stage-Gate process: What best-practice companies do—I, *Research-Technology Management*, Vol. 45, No. 5, 21-27, doi: 10.1080/08956308.2002.11671518.

[16] Kruachottikul, P., Dumrongvute, P., Tea-makorn, P., Kittikowit, S., Amrapala, A. (2023). New product development process and case studies for deep-tech academic research to commercialization, *Journal of Innovation and Entrepreneurship*, Vol. 12, No. 1, Article No. 48, doi: 10.1186/s13731-023-00311-1.

[17] Prado Saldanha, F., Cohendet, P., Pozzebon, M. (2014). Challenging the Stage-Gate model in crowdsourcing: The case of Fiat Mio in Brazil, *Technology Innovation Management Review*, Vol. 4, No. 9, 28-35, doi: 10.22215/timreview/829.

[18] Drucker, P.F. (1985). *Innovation and entrepreneurship: Practice and principles*, Harper & Row, New York, USA.

[19] Christensen, C.M. (2013). *Innovator's solution: Creating and sustaining successful growth*, Harvard Business Press, New York, USA.

[20] Li, Q., Zheng, G. (2008). How to win competitive advantage through disruptive innovation, In: *Proceedings of Annual Conference of the Academy of Innovation and Entrepreneurship*, Beijing, China, 78-81.

[21] Pretnar, B. (2020). *Intelektualna lastnina in tržno uspešne inovacije: Priročnik za managerje, raziskovalce in izumitelje*, Lexpera, GV založba, Ljubljana, Slovenia.

[22] Chen, Y.K., Coviello, N., Ranaweera, C. (2021). When change is all around: How dynamic network capability and generative NPD learning shape a firm's capacity for major innovation, *Journal of Product Innovation Management*, Vol. 38, No. 5, 574-599, doi: 10.1111/jpim.12595.

[23] Cooper, R.G., Edgett, S.J., Kleinschmidt, E.J. (2004). Benchmarking best NPD practices—III, *Research-Technology Management*, Vol. 47, No. 6, 43-55, doi: 10.1080/08956308.2004.11671662.

[24] Kruger, L.L.S.J., Pretorius, J H.C., Erasmus, L.D. (2019). Towards a comprehensive systematic innovation model: A literature review, *SAIEE Africa Research Journal*, Vol. 110, No. 1, 39-46, doi: 10.23919/saiee.2019.8643149.

[25] Tidd, J., Bessant, J.R. (2020). *Managing innovation: Integrating technological, market and organizational change*, Wiley, Chichester, UK.

[26] Gärtner, Q., Ronco, E., Cagliano, A.C., Reinhart, G. (2023). Development of an approach for the holistic assessment of innovation projects in manufacturing including potential, effort, and risk using a systematic literature review and expert interviews, *Applied Sciences*, Vol. 13, No. 5, Article No. 3221, doi: 10.3390/app13053221.

[27] Schuhmacher, A., Gassmann, O., Hinder, M. (2016). Changing R&D models in research-based pharmaceutical companies, *Journal of Translational Medicine*, Vol. 14, No. 1, Article No. 105, doi: 10.1186/s12967-016-0838-4.

[28] Cooper, R.G., Edgett, S.J., Kleinschmidt, E.J. (2004). Benchmarking best NPD practices—II, *Research-Technology Management*, Vol. 47, No. 3, 50-59, doi: 10.1080/08956308.2004.11671630.

[29] Kahn, K.B. (2018). Understanding innovation, *Business Horizons*, Vol. 61, No. 3, 453-460, doi: 10.1016/j.bushor.2018.01.011.

[30] Griffin, A. (1997). PDMA research on new product development practices: Updating trends and benchmarking best practices, *Journal of Product Innovation Management*, Vol. 14, No. 6, 429-458, doi: 10.1111/1540-5885.1460429.

[31] Markides, C. (2006). Disruptive innovation: In need of better theory, *Journal of Product Innovation Management*, Vol. 23, No. 1, 19-25, doi: 10.1111/j.1540-5885.2005.00177.x.

[32] Christensen, C.M., Anthony, S.D., Roth, E.A. (2005). *Korak pred prihodnostjo: Kako s teorijami o inovacijah napovedati spremembe v industriji*, GV založba, Ljubljana, Slovenia.

[33] Moore, G.A. (2006). *Crossing the chasm: Marketing and selling disruptive products to mainstream customers*, Harper Business Essentials, New York, USA.

[34] Pegkas, P., Staikouras, C., Tsamadias, C. (2019). Does research and development expenditure impact innovation? Evidence from the European Union countries, *Journal of Policy Modeling*, Vol. 41, No. 5, 1005-1025, doi: 10.1016/j.jpolmod.2019.07.001.

[35] Cooper, R.G., Edgett, S.J., Kleinschmidt, E.J. (2004). Benchmarking best NPD practices—I, *Research-Technology Management*, Vol. 47, No. 1, 31-43, doi: 10.1080/08956308.2004.11671606.

[36] Bianchi, M., Marzi, G., Guerini, M. (2020). Agile, Stage-Gate and their combination: Exploring how they relate to performance in software development, *Journal of Business Research*, Vol. 110, 538-553, doi: 10.1016/j.jbusres.2018.05.003.

[37] Cooper, R.G. (1985). Selecting winning new product projects: Using the NewProd system, *Journal of Product Innovation Management*, Vol. 2, No. 1, 34-44, doi: 10.1111/1540-5885.210034.

[38] Sommer, A.F., Hedegaard, C., Dukovska-Popovska, I., Steger-Jensen, K. (2015). Improved product development performance through agile/Stage-Gate hybrids: The next-generation Stage-Gate process?, *Research-Technology Management*, Vol. 58, No. 1, 34-45, doi: 10.5437/08956308x5801236.

[39] Walrave, B., Dolmans, S.A.M., van Oorschot, K.E., Nuijten, A.L.P., Keil, M., van Hellemond, S. (2022). Dysfunctional agile–stage-gate hybrid development: Keeping up appearances, *International Journal of Innovation and Technology Management*, doi: 10.1142/S0219877022400041.

[40] Cooper, R.G., Sommer, A.F. (2016). Agile-Stage-Gate: New idea-to-launch method for manufactured new products is faster, more responsive, *Industrial Marketing Management*, Vol. 59, 167-180, doi: 10.1016/j.indmarman.2016.10.006.

[41] Grönlund, J., Sjödin, D.R., Frishammar, J. (2010). Open innovation and the Stage-Gate process: A revised model for new product development, *California Management Review*, Vol. 52, No. 3, 106-131, doi: 10.1525/cmr.2010.52.3.106.

[42] Cooper, R.G. (1983). A process model for industrial new product development, *IEEE Transactions on Engineering Management*, Vol. EM-30, No. 1, 2-11, doi: 10.1109/tem.1983.6448637.

[43] Cooper, R.G. (1999). The invisible success factors in product innovation, *Journal of Product Innovation Management*, Vol. 16, No. 2, 115-133, doi: 10.1111/1540-5885.1620115.

[44] Milenkovic, M., Ciric Lalic, D., Vujicic, M., Pesko, I., Savkovic, M., Gracanin, D. (2023). Project portfolio management in telecommunication company: A stage-gate approach for effective portfolio governance, *Advances in Production Engineering & Management*, Vol. 18, No. 3, 357-370, doi: 10.14743/apem2023.3.478.

[45] Cooper, R.G. (2014). What's next?: After Stage-Gate, *Research-Technology Management*, Vol. 57, No. 1, 20-31, doi: 10.5437/08956308x5606963.

[46] Cooper, R.G., Sommer, A.F. (2016). The agile-Stage-Gate hybrid model: A promising new approach and a new research opportunity, *Journal of Product Innovation Management*, Vol. 33, No. 5, 513-526, doi: 10.1111/jpim.12314.

[47] Rochford, L., Rudelius, W. (1997). New product development process: Stages and successes in the medical products industry, *Industrial Marketing Management*, Vol. 26, No. 1, 67-84, doi: 10.1016/s0019-8501(96)00115-0.

[48] Schmidt, S., von der Oelsnitz, D. (2020). Innovative business development: Identifying and supporting future radical innovators, *Leadership, Education, Personality: An Interdisciplinary Journal*, Vol. 2, No. 1, 9-21, doi: 10.1365/s42681-020-00008-z.

[49] OECD, Eurostat (2018). *Oslo Manual 2018*, OECD Publishing, Paris, France, doi: 10.1787/9789264304604-en.

[50] Eling, K., Langerak, F., Griffin, A. (2013). A stage-wise approach to exploring performance effects of cycle time reduction, *Journal of Product Innovation Management*, Vol. 30, No. 4, 626-641, doi: 10.1111/jpim.12019.

[51] Christensen, C.M. (2013). *The innovator's dilemma: When new technologies cause great firms to fail*, Harvard Business Review Press, New York, USA.

[52] Cooper, R., Mills, M. (2005). How P&G achieves such stellar NPD results, *PDMA Visions*.

[53] Allen, G.J. (2022). Conceptualize™: A new contribution to generate real-needs-focussed, user-centred, lean business models, *Journal of Innovation and Entrepreneurship*, Vol. 11, No. 1, Article No. 6, doi: 10.1186/s13731-022-00198-4.

[54] Lewrick, M., Link, P., Leifer, L. (2020). *The design thinking toolbox: A guide to mastering the most popular and valuable innovation methods*, Wiley, Hoboken, USA.

[55] Nakata, C., Hwang, J. (2020). Design thinking for innovation: Composition, consequence, and contingency, *Journal of Business Research*, Vol. 118, 117-128, doi: 10.1016/j.jbusres.2020.06.038.

[56] Franchini, G., Dosi, C., Vignoli, M. (2017). The coexistence of design thinking and stage and gate in the same organisational context - Challenges and need for integration, In: *Proceedings of the 21st International Conference on Engineering Design (ICED 17)*, Vancouver, Canada, 301-310.

[57] DelVecchio, J., White, F., Phelan, S.E. (2013). Tools for innovation management: A comparison of lean startup and the Stage-Gate system, *SSRN*, doi: 10.2139/ssrn.2534138.

[58] Maurya, A. (2016). *Delaj vitko: Od načrta A do načrta, ki deluje*, Pasadena, Ljubljana, Slovenia.

[59] Eddoug, F.-Z., Benabbou, R., Benhra, J. (2023). Adapting P2M framework for innovation program management through a lean-agile approach, *International Journal of Information Technology Project Management*, Vol. 14, No. 1, 1-18, doi: 10.4018/ijitpm.318125.

[60] Nicoletti, B. (2015). Optimizing innovation with the lean and digitize innovation process, *Technology Innovation Management Review*, Vol. 5, No. 3, 29-38, doi: 10.22215/timreview/879.

[61] van der Duin, P.A., Ortt, J.R., Aarts, W.T.M. (2014). Contextual innovation management using a Stage-Gate platform: The case of Philips shaving and beauty, *Journal of Product Innovation Management*, Vol. 31, No. 3, 489-500, doi: 10.1111/jpim.12109.

[62] Goldasteh, P., Akbari, M., Bagheri, A., Mobini, A. (2022). How high-tech start-ups learn to cross the market chasm?, *Journal of Global Entrepreneurship Research*, Vol. 12, 157-173, doi: 10.1007/s40497-022-00316-2.

[63] Gadd, K., Goddard, C. (2011). *TRIZ for engineers: Enabling inventive problem solving*, Wiley, Chichester, UK.

[64] Abramov, O.Y. (2014). TRIZ-assisted Stage-Gate process for developing new products, *Journal of Finance and Economics*, Vol. 2, No. 5, 178-184, doi: 10.12691/jfe-2-5-8.

[65] Abramov, O.Y., Medvedev, A.V., Rychagov, V.Y. (2019). Evaluation of the effectiveness of modern TRIZ based on practical results in new product development, In: Benmoussa, R., De Guio, R., Dubois, S., Koziołek, S. (eds.), *New Opportunities for Innovation Breakthroughs for Developing Countries and Emerging Economies. TFC 2019. IFIP Advances in Information and Communication Technology,* Vol 572, Springer, Cham. 43-52, doi: 10.1007/978-3-030-32497-1_4.

[66] Louw, L., Schutte, C.S.L., Seidel, C., Imser, C. (2018). Towards a flexible innovation process model assuring quality and customer needs, *South African Journal of Industrial Engineering*, Vol. 29, No. 1, 155-168, doi: 10.7166/29-1-1911.

[67] Abramov, O., Kogan, S., Mitnik-Gankin, L., Sigalovsky, I., Smirnov, A. (2015). TRIZ-based approach for accelerating innovation in chemical engineering, *Chemical Engineering Research and Design*, Vol. 103, 25-31, doi: 10.1016/j.cherd.2015.06.012.

[68] Rantanen, K., Domb, E. (2007). *Simplified TRIZ: New problem solving applications for engineers and manufacturing professionals*, Second Edition, Taylor & Francis, Boca Raton, USA, doi: 10.1201/9781420000320.

[69] Trebar, A. (2010). *Kreativno snovanje novih proizvodov in storitev,* Slovensko združenje za kakovost in odličnost, Ljubljana, Slovenia.

[70] Orloff, M.A. (2018). *ABC-TRIZ: Introduction to creative design thinking with modern TRIZ modeling*, Springer International Publishing, Cham, Switzerland.

[71] Markham, S.K. (2013). The impact of front-end innovation activities on product performance, *Journal of Product Innovation Management*, Vol. 30, No. s1, 77-92, doi: 10.1111/jpim.12065.

[72] Livotov, P. (2016). Estimation of new-product success by company's internal experts in the early phases of innovation process, *Procedia CIRP*, Vol. 39, 150-155, doi: 10.1016/j.procir.2016.01.181.

[73] Kock, A., Heising, W., Gemünden, H.G. (2014). How ideation portfolio management influences front-end success, *Journal of Product Innovation Management*, Vol. 32, No. 4, 539-555, doi: 10.1111/jpim.12217.

[74] Eling, K., Langerak, F., Griffin, A. (2015). The performance effects of combining rationality and intuition in making early new product idea evaluation decisions, *Creativity and Innovation Management*, Vol. 24, No. 3, 464-477, doi: 10.1111/caim.12128.

[75] Koen, P. A., Bertels, H.M.J., Kleinschmidt, E. (2014). Managing the front end of innovation—Part I: Results from a three-year study, *Research-Technology Management*, Vol. 57, No. 2, 34-43, doi: 10.5437/08956308x5702145.

[76] Cooper, R.G. (2010). The Stage-Gate idea to launch system, *Wiley International Encyclopedia of Marketing*, Wiley, Chichester, UK, doi: 10.1002/9781444316568.wiem05014.

[77] Wojciechowski, A., Becker, B., Kirchner, M., Kreidler, B. (2019). Implementation of sustainability in innovation management: The Idea to People, Planet and Profit (I2P^3®) process, *Journal of Business Chemistry*, Vol. 16, No. 1, 26-40.

[78] Wen, Y., Li, X., Yan, R., Zhu, X. (2024). A comparative empirical study on the product R&D capability of manufacturing industry, *Tehnički Vjesnik – Technical Gazette,* Vol. 31, No. 1, 88-97.

[79] van de Burgwal, L.H.M., Ribeiro, C.D.S., van der Waal, M.B., Claassen, E. (2018). Towards improved process efficiency in vaccine innovation: The vaccine innovation cycle as a validated, conceptual stage-gate model, *Vaccine*, Vol. 36, No. 49, 7496-7508, doi: 10.1016/j.vaccine.2018.10.061.

[80] Sun, D., Gao, W., Hu, H., Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it?, *Acta Pharmaceutica Sinica B*, Vol. 12, No. 7, 3049-3062, doi: 10.1016/j.apsb.2022.02.002.

[81] Pervaz, J., Sremčev, N., Stevanov, B., Gusel, L. (2024). Simulation-based algorithm for continuous improvement of enterprises performance, *International Journal of Simulation Modelling*, Vol. 23, No. 2, 215-226, doi: 10.2507/IJSIMM23-2-670.

[82] Hackl, C., Guillermin, S. (2020). Integration of IP into the 'classical' Stage-Gate model, *les Nouvelles - Journal of the Licensing Executives Society*, Vol. 55, No. 2, 143-147.

[83] Kumar, M., Beninger, S., Reppel, A., Stanton, J., Vlaminck, D., Watson, F. (2025). Your synthetic teammate: Enriching new product development with generative AI, *Business Horizons*, In press, Journal pre-proof, doi: 10.1016/j.bushor.2025.02.008.

[84] He, Y., Song, J., Ouyang, W. (2024). Digital economy, entrepreneurship, and high-quality development of the manufacturing industry, *Tehnički Vjesnik – Technical Gazette,* Vol. 31, No. 3, 851-863, doi: 10.17559/TV-20231121001135.

[85] Li, J., Yang, F., Qi, J., Sun, R., Geng, R. (2022). The influence of job satisfaction on entrepreneurial intention: A cross-level investigation, *International Small Business Journal*, Vol. 40, No. 3, 385-402, doi: 10.1177/02662426211018831.

[86] Indrawati, H., Caska, C., Suarman, S. (2020). Barriers to technological innovations of SMEs: How to solve them?, *International Journal of Innovation Science*, Vol. 12, No. 5, 545-564, doi: 10.1108/ijis-04-2020-0049.

[87] Stojic, N., Delic, M., Bojanic, T., Jokanovic, B., Tasic, N. (2024). Integrated model of risk management in business processes in industrial systems, *International Journal of Simulation Modelling*, Vol. 23, No. 3, 412-423, doi: 10.2507/IJSIMM23-3-689.

Calendar of events

- International Conference on Smart Materials & Structures, March 27-28, 2025, Berlin, Germany.
- 14th Spring World Congress on Engineering and Technology (SCET 2025), April 19-21, 2025, Guilin, China.
- International Connect & Expo on Material Science and Engineering, April 28-20, 2025, Rome, Italy.
- 2025 Annual Modeling and Simulation Conference (ANNSIM'25), May 26-29, 2025, Madrid, Spain.
- Fifth International Conference on Simulation for Additive Manufacturing (Sim-AM 2025), September 9-11, 2025, Pavia, Italy.

Notes for contributors

General

Articles submitted to the *APEM journal* should be original and unpublished contributions and should not be under consideration for any other publication at the same time. Manuscript should be written in English. Responsibility for the contents of the paper rests upon the authors and not upon the editors or the publisher. The content from published paper in the *APEM journal* may be used under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0). For most up-to-date information please see the APEM journal homepage *apem-journal.org*.

Submission of papers

A submission must include the corresponding author's complete name, affiliation, address, phone and fax numbers, and e-mail address. All papers for consideration by *Advances in Production Engineering & Management* should be submitted by e-mail to the journal Editor-in-Chief:

Miran Brezocnik, Editor-in-Chief
UNIVERSITY OF MARIBOR
Faculty of Mechanical Engineering
Chair of Production Engineering
Smetanova ulica 17, SI – 2000 Maribor
Slovenia, European Union
E-mail: apem@apem-journal.org

Manuscript preparation

Manuscript should be prepared in *Microsoft Word 2010* (or higher version) word processor. *Word .docx* format is required. Papers on A4 format, single-spaced, typed in one column, using body text font size of 11 pt, should not exceed 12 pages, including abstract, keywords, body text, figures, tables, acknowledgements (if any), references, and appendices (if any). The title of the paper, authors' names, affiliations and headings of the body text should be in *Calibri* font. Body text, figures and tables captions have to be written in *Cambria* font. Mathematical equations and expressions must be set in *Microsoft Word Equation Editor* and written in *Cambria Math* font. For detail instructions on manuscript preparation please see instruction for authors in the *APEM journal* homepage *apem-journal.org*.

The review process

Every manuscript submitted for possible publication in the *APEM journal* is first briefly reviewed by the editor for general suitability for the journal. Notification of successful submission is sent. After initial screening, and checking by a special plagiarism detection tool, the manuscript is passed on to at least two referees. A double-blind peer review process ensures the content's validity and relevance. Optionally, authors are invited to suggest up to three well-respected experts in the field discussed in the article who might act as reviewers. The review process can take up to eight weeks on average. Based on the comments of the referees, the editor will take a decision about the paper. The following decisions can be made: accepting the paper, reconsidering the paper after changes, or rejecting the paper. Accepted papers may not be offered elsewhere for publication. The editor may, in some circumstances, vary this process at his discretion.

Proofs

Proofs will be sent to the corresponding author and should be returned within 3 days of receipt. Corrections should be restricted to typesetting errors and minor changes.

Offprints

An e-offprint, i.e., a PDF version of the published article, will be sent by e-mail to the corresponding author. Additionally, one complete copy of the journal will be sent free of charge to the corresponding author of the published article.

APEM journal

Advances in Production Engineering & Management

Chair of Production Engineering (CPE)
University of Maribor
APEM homepage: apem-journal.org

Volume 20 | Number 1 | March 2025 | pp 1-152

Contents

Published by CPE, University of Maribor

ISSN 1854-6250
9 771854 625008

apem-journal.org