J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 4913--4929

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on the multi-objective operation efficiency improvement path of enterprises based on reinforcement learning algorithm

Jinhua Ma[1,✉], Rong Zhu[2]

[1] *Department of Economics and Management, Shandong Vocational College of Science and Technology, Weifang, Shandong, 261000, China*

[2] *College of Economics and Management, Qingdao Institute of Technology, Qingdao, Shandong, 266300, China*

ABSTRACT

In enterprise operations, multi-objective optimization involves multiple conflicting objectives such as cost escalation control, customer satisfaction, and production efficiency. Based on reinforcement learning algorithm, the article deals with multi-objective optimization problem in enterprise operation through the interactive learning between intelligent body and environment, for which a multi-objective operation efficiency improvement path for enterprise based on Q-learning scheduling is designed. The simulation data is utilized to generate the PDR tree structure, and subsequently, the intelligent body is prompted to complete the multi-objective operation learning of the enterprise through several iterations. On this basis, the intelligent body completes all the actions and generates scheduling strategies to improve operational efficiency. The model proposed in this paper can predict the demand changes of enterprises in the future time window and make the best decision to improve the operational efficiency. Under the model of this paper, the mean values of pure technical efficiency as well as scale efficiency of 10 firms in 2024 are 0.9 and 0.933, respectively, and they are predicted to continue to grow in 2025. The model reduces the firms' average operating costs and administrative expenses, while employee compensation and fixed assets increase by 49.58% and 19.48%. Since the survey period, the TFP index of all 10 companies is greater than 1, which indicates that, the application of the model in this paper improves the operational efficiency of the companies.

Keywords: Reinforcement learning; Q-learning algorithm; PDR tree; scheduling strategy; operational efficiency

✉ Corresponding author.
E-mail address: jinhuama@sdvcst-edu.cn (J. Ma).
Received 03 February 2024; Revised 12 April 2024; Accepted 10 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-277

1. Introduction

Reinforcement learning is an algorithm in the field of machine learning that learns optimal behavior through interaction with the environment. The core idea of a reinforcement learning algorithm is to learn by trial and error, constantly adjusting one's strategy to maximize the cumulative rewards of learning. Reinforcement learning algorithms have a wide range of applications in many fields, such as games, robot control, financial transactions, etc. It is important to use them as a path to improve the operational efficiency of enterprises with multiple objectives [1-4].

Improving the operational efficiency of an enterprise is an issue that every enterprise attaches great importance to. By improving operational efficiency, enterprises can increase productivity, reduce costs, enhance competitiveness, and realize higher profitability. And mastering appropriate strategies can effectively improve the operational efficiency of enterprises [5-8]. First of all, an organization needs to have clear and explicit goals and strategies to improve operational efficiency. This can help the organizational team work better together and make the work targeted and efficient [9-11]. Secondly, a good organizational structure can facilitate the flow of information, rapid implementation of decision-making and so on. Optimizing the organizational structure, rationally dividing responsibilities and authorities, coordinating work among departments, and reducing redundancy and duplication of work. At the same time, review and optimize the business processes, remove inefficient links, and improve efficiency [12-15]. In addition, employees are the most important assets of the enterprise, and their ability directly affects the operational efficiency of the enterprise. Therefore, emphasizing on employee training and enhancement is one of the keys to improve the operational efficiency of an enterprise. Continuous training programs can help employees learn new skills and knowledge, and improve work efficiency and quality [16-19]. In addition, an efficient enterprise cannot be separated from good communication and collaboration mechanisms. Good communication can avoid distortion and misunderstanding of information transmission and improve work efficiency. Meanwhile, encouraging collaboration and cooperation among team members can promote the sharing of knowledge and experience, and solve problems more quickly and efficiently [20-23].

Literature [24] proposed the concept of evaluation and improvement of the operational efficiency of enterprise quality management systems. The problem of developing a theoretical tool for evaluating and managing the operational efficiency of quality management systems as a theoretical tool for sustainable development of enterprises was solved. The aim of the literature [25] is to test an innovative strategy for improving the efficiency of enterprise management to reduce waste and the formation of manufacturing defects. The results can be targeted to influence the increase in the efficiency of enterprise management through innovative management approaches. Literature [26] examined the impact of standardization of logistics services on the operational efficiency, performance and investment activities of firms. It revealed that standardization facilitates the improvement of operational efficiency of firms, which is important for many types of firms and in regions with high marketization. Literature [27] assessed the operational efficiency of e-commerce companies in China and pointed out that the overall efficiency, business capability and business scale are in a state of gradual improvement. It also puts forward suggestions such as adjusting scale and enhancing competitive advantages. Literature [28] explores the impact of government subsidies on the operational efficiency of enterprises and the characteristics of zombie enterprises that are “dead or dying”. It shows that although subsidies improve the operational efficiency of enterprises to a certain extent, they are less effective in improving the efficiency and profitability of zombie enterprises. This finding helps to improve the subsidy policy. Literature [29] shows that by improving communication within firms helps to improve the operational efficiency of firms. It also describes that firms design ERP in order to improve the operational efficiency of the manufacturing system in order to reduce work and inventory. Literature [30] explored the potential impact of DT on the operational efficiency

(OE) of electric utilities. It shows that DT indirectly positively affects the OE of electric utilities by promoting innovation, improving capital utilization and alleviating financial constraints. These effects are more significant in a competitive industry environment. Literature [31] examines the impact of digital transformation on firms. It is indicated that the digital transformation of enterprises contributes to the efficiency of their technological innovation. This impact is heterogeneous and is reflected in both factor intensity and nature of ownership. Literature [32] used DEA data envelopment analysis to compare the operational efficiency of companies implementing green M&As in China. Conclusions such as the uneven distribution of green M&A firms in the eastern, central and western regions were drawn, and suggestions were made to the sustainable development of heavily polluting firms. From the above study, it is easy to see that the improvement of the operational efficiency of enterprises can be realized from the improvement of internal communication, government subsidies, and enterprise innovation, and emphasizes the importance of the improvement of the operational efficiency of enterprises for the sustainable development of enterprises. And reinforcement learning as a machine learning method, its application can bring optimal decision-making for enterprises, but as of now, the application in enterprise operation efficiency has not been paid attention to.

This paper discusses the design process and application study of multi-objective operational efficiency improvement path for enterprises based on reinforcement learning algorithms. The study uses binary tree to represent PDR parameters to generate scheduling strategies, and leaf nodes containing operations are generated through iteration. Subsequently, a PDR library is constructed to optimize the best performing PDRs to improve the scheduling problems in different operations. Using the Q-learning algorithm, the intelligences are trained to make actions to improve the efficiency of business operations and get the corresponding rewards as a path to improvement. The robustness of the efficiency improvement performance is verified in the data collected in this paper, and the effectiveness of the path proposed in this paper is evaluated in combination with the results of practical applications in several enterprises.

2. Key technologies and methodologies

2.1. Definition of the research base and measurement methods

2.1.1. Operational efficiency of enterprises. Efficiency is the input-output ratio, the decision-making unit in the allocation of resources are expected to obtain more output with less input. Enterprise operational efficiency [33] can be defined as the degree to which the enterprise realizes its business objectives, mainly refers to the relationship between various input and output elements in the process of enterprise operation. In the context of market economy, enterprise operational efficiency can not only assess whether the enterprise's resource allocation is reasonable, but also reflect the enterprise's profitability, and then seek to improve the direction. In order to pursue higher profits, enterprises need to continuously carry out reform and innovation, adjust their own operation and management mode, in order to improve operational efficiency, and thus enhance their competitiveness in the market and obtain more development opportunities.

For listed companies, according to the production frontier theory, the operational efficiency of enterprises can be divided into three parts for evaluation, namely, integrated efficiency (OTE), pure technical efficiency (PTE) and scale efficiency (SE). Among them, comprehensive efficiency can be decomposed into pure technical efficiency and scale efficiency, which is the product of the two, and can measure the resource allocation and resource utilization efficiency of listed companies at this stage, reflecting their production and operation status. Pure technical efficiency can be used to evaluate the impact of changes in input factors such as management and technology on the efficiency of the enterprise, when the listed company's management level or production technology lags behind, the efficiency of the enterprise's operations will decline, and even limit the development of the enterprise. Scale efficiency, on the other hand, reflects the impact of changes in the enterprise's own scale on its production and operation process; expanding investment in infrastructure construction, purchasing production equipment and other ways can improve the scale efficiency of the enterprise.

2.1.2. Measurement methods:

1) DuPont Financial Analysis System. DuPont analysis method is to take the key financial indicator return on net assets of an enterprise as a starting point, and through the splitting of this indicator and the introduction of other relevant financial indicators, to conduct an integrated and comprehensive analysis of the enterprise's financial position, so as to map out the enterprise's operational efficiency. It is a method to measure the operational efficiency of an enterprise by utilizing the data related to its operational efficiency and return on shareholders' equity.

2) Enterprise Credit Rating Method. Credit rating mainly refers to the activities of credit rating agencies in assessing the credit situation of an enterprise on a hierarchical basis by collecting relevant credit information of the target enterprise and utilizing certain credit indicators. Credit rating covers a wide range, involving many real enterprises and financial enterprises, which analyzes the financial situation of the enterprise at the same time, but also the enterprise's profitability, internal management, enterprise quality and industry, as well as the risks faced by the enterprise are incorporated into the evaluation system, which can reflect the overall situation of an enterprise in a more comprehensive manner, and the credit evaluation system of insurance enterprises is such a complex but comprehensive integrated system.

3) Three-stage DEA modeling method. Data Envelopment Analysis [34], as an effective business efficiency analysis model, is based on linear programming and convex analysis, and compares and evaluates the efficiency between different decision-making units through the use of mathematical planning models. In detail, the data envelopment analysis method begins by substituting multiple input-output decision-making units into the original DEA model to measure the efficiency values under the traditional model. Then SFA analysis is applied to implement regression analysis on slack variables and environmental factors, and remove the interference of environmental variables: finally, the optimized variables and outputs are substituted into the DEA model, and another quantitative analysis is conducted to obtain the accurate efficiency value with the removal of environmental variables and other influencing factors. Because the DEA method synthesizes the data in an indirect way, the data do not need to be dimensionless again, so it has an absolute advantage in studying the efficiency of multiple inputs and multiple outputs.

2.2. Enhanced Learning Model

2.2.1. Enhanced learning system. Reinforcement learning [35], as one of the techniques in the field of artificial intelligence, enables intelligences to deal with decision-making problems by learning through interaction with the environment. Intelligent bodies are able to gain knowledge about the environment during continuous interactive learning and use the existing experience to gain rewards and choose to perform actions while exploring, allowing for a better space of action choices in the future. The intelligences use this knowledge to extend past learning experiences to new environments, thus being generalizable. Figure 1 shows the system framework of the reinforcement learning algorithm.

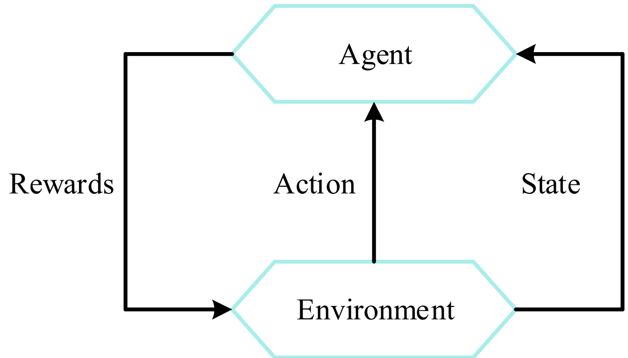

Fig. 1. Reinforcement learning system framework

2.2.2. Markov decision-making process. Simply put, Markov Decision Process [36] is a continuously repeated learning process in which an intelligent body changes its current state by executing actions to obtain rewards and interacts with its environment. The Markov property can be reflected in the learning process, specifically, the MDP depends only on the current state, independent of subsequent states, actions, etc. The MDP consists of a quintuple $< S, A, P, R, \gamma >$ as follows:

1) $s \in S : S$ represents a finite set of all states, and s represents the current state at a given moment.

2) $a \in A : A$ represents a finite set of all actions, and a represents the current action at a certain moment.

3) State transfer function: $P(S, A, S^{'}) \sim P(S_{i+1} = s^{'} \mid S_i = s, A_i = a)$ represents the probability of taking action a from the current state s to the next state $s^{'}$.

4) Reward function: $R(s,a) = E[R_{t+1} \mid S_t = s, A_t = a]$ represents the reward reward of an intelligent body after taking an action at a certain moment.

5) Discount Factor $\gamma \in [0,1]$: In order to balance immediate and delayed rewards, reinforcement learning uses Discount Factor γ to calculate the cumulative reward value over time.

Reinforcement learning wants to focus more on immediate rewards, and for delayed rewards introduces discount factor γ for attenuation, thus converting an infinite time series problem into a problem with a maximum upper bound, and equation (1) defines the cumulative rewards U. The purpose of reinforcement learning is to maximize future long-term rewards, i.e., to find the maximum U, which implies that the rewards of the present moment are more important than those of the future feedback, and this is also fits the reality of human brain learning:

$$U_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T \tag{1}$$

An action a selected by an intelligent body at the current moment t during the learning process is executed based on the strategy π, which represents the correspondence between the state of the intelligent body at this moment and the action with the following expression:

$$\pi(a \mid s) = P[A_t = a \mid S_t = s] \tag{2}$$

The state value function is the expectation of the cumulative return that can be obtained in state s at moment t based on strategy π, and the formula can be expressed as follows:

$$V_\pi(s) = R_{t+1} + R_{t+2} + \cdots = E_\pi[U_t \mid S_t = s] \tag{3}$$

The action value function is the expectation of the cumulative payoff that can be obtained at t moment based on strategy π, state s after choosing an action a, and the formula can be expressed as follows:

$$Q_\pi(s,a) = E_\pi[U_t \mid S_t = s, A_t = a] \tag{4}$$

2.3. Improvement path of enterprise operation efficiency based on Q-learning scheduling

2.3.1. Constructing the PDR tree. In this paper, the process of PDR (Priority Scheduling Rules) generating scheduling policies according to parameters is represented as a binary tree. All the generation optimized operations are the parameters of the selected PDR with $X_1 = \{job_1, job_2, ..., job_n\}$. P, and the size of the X_1 parameter values P_l is bounded into sets of operations X_2 and X_3, and the solution formula for P_l is shown in Equation (5):

$$P_l = \frac{1}{d}\sum_{k=1}^{card(X_l)} p_{lk} \quad l \in \{1, 2, \cdots, L\} \tag{5}$$

where d denotes a hyperparameter used to control the direction of tree growth, $card(X_l)$ denotes the number of businesses contained in the tree node, and p_k denotes the value of the parameter for the kth business in tree node l. Iterate sequentially until a leaf node containing only one business is finally generated.

2.3.2. PDR Preferences. The performance of different PDRs in different scheduling problems varies greatly, it is necessary to build a PDR library that covers a variety of initial priority scheduling rules, and optimize the PDRs with strong performance. On this basis, further improvement is made to get the PDRs with more excellent performance in the end. In this paper, the parameter of the PDRs is chosen to be the time for the machine to complete the service, and the representation of the PDR library is shown in Equation (6):

$$PDR\_set = \{SPT_i, STPT \mid i = 1, 2, \cdots, 6\} \tag{6}$$

Where SPT_i denotes that the shorter the processing time of a service at machine i the more it is prioritized to be processed. STPT denotes that the shorter the total processing time of a service the more it is prioritized to be processed.

2.3.3. Improvement of the PDR tree. There are four important concepts in Reinforcement Learning, which are Intelligent Body, Reward, Strategy and Environment.

Environment: the intelligent body can perceive the current state of the environment, and after its own calculation gives the action of the current round, which will be acted into the environment, and the environment then sends out the reward signal and produces the corresponding state transfer, which is expressed by the formula:

$$\text{Next moment state} \sim \text{P}(\cdot \mid \text{current state, agent action}) \tag{7}$$

Intelligent Body: An intelligent body is a machine that can make decisions by sensing information about its surroundings, or it can directly change this environment based on its decisions. In each iteration, the intelligent body can sense the current state of the environment s_t, and through computation give the current round of action A_t, which acts on the environment, and the environment then gives the corresponding reward R_t and state transfer S_{t+1}. In the next round of interaction, the intelligences can perceive the new environment state and repeat the iteration.

Reward: the environment generates a scalar signal as a reward feedback based on the state and the action taken by the intelligent body, which measures how good or bad the intelligent body's action is in this round, denoted as R.

Strategy: the process by which an intelligent body computes the actions it needs to take to achieve a goal based on its current state is called decision making, denoted as π. Strategy is the final form of intelligence embodied by an intelligent body, and it is the core difference between different intelligences.

In this paper, the environment is the already constructed PDR tree, and when the two children of the parent node are swapped, the state of the environment is updated from S_t to S_{t+1}. The state of the environment is the new tree structure generated by the intelligences after they have made an action, and each tree structure is a new PDR.

The intelligent body is the method for choosing different actions, i.e., deciding whether to swap the child nodes or not.

Actions: In this paper, there are only two types of actions that an intelligent body can make, i.e., replacing the two children of the current node, denoted as $a=1$. And keeping the position of the two children, denoted as $a=0$. The direction of making an action is from top to bottom and from left to right. So the process of transferring the state of the environment is also from the root node to the leaf nodes of the tree, from the left node to the right node.

Reward: the difference of the maximum completion time of the current state relative to the PDR of the previous state as shown in equation (8):

$$r_t = pdr^{t}_{makespan}(X) - pdr^{t+1}_{makespan}(X) \quad t \in 1,2,\cdots,T-1 \tag{8}$$

$pdr^{t}_{makespan}(X)$ denotes the maximum completion time calculated from the PDR in the current state. In this round of iteration, in addition to the initial PDR tree, there are $T-1$ states, $T-1$ tree structures, and X represents the matrix consisting of the processing time of the production system operations on the machine.

Fig. 2 shows the learning process of an intelligent body at one time, in the environment state s_0, i.e., the initial state the intelligent body decides whether the action $a_1 (a_1 \in \{0,1\})$ exchanges the two children of the current node or not, obtains the reward r_1, and the environment is updated to s_i. Under the condition of environment state s_i, the intelligent body makes the action a_2, obtains the reward r_2, and the environment is updated to s_2. And so on, completes all the actions and obtains the final reward.

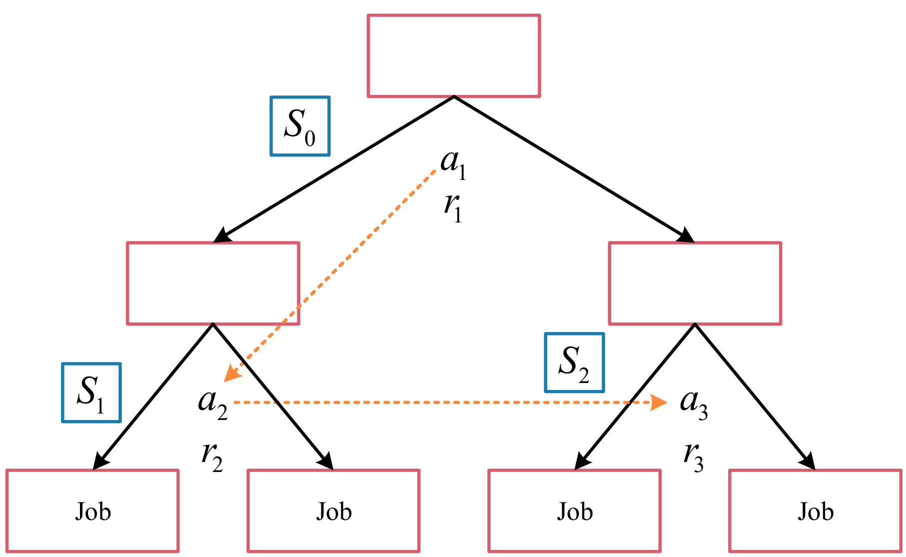

Fig. 2. A learning process of the agent

2.3.4. Q-Learning Algorithm. The Q-Learning algorithm [37] uses the ε-Greedy algorithm for searching and learning, where searching means trying a new action with a certain probability ε. Learning refers to making the action that maximizes the expected reward valuation from previous experience with a probability of $1-\varepsilon$. The combination of the two allows the intelligent body to make appropriate attempts to find a better, but will eventually return to a better result. The formula is shown in (9):

$$a_t = \begin{cases} \arg\max_{a \in A} \hat{Q}(a), & \text{Sampling frequency} : 1-\varepsilon \\ \text{Choose at random from } A & \text{Sampling probability} : \varepsilon \end{cases} \tag{9}$$

A is the set of all selectable actions in this decision.

The merit of a reinforcement learning strategy can be represented by the action value function $Q(s_i, a_i)$, i.e., the value of action a_i conditional on the state s_i of the environment. The temporal difference method is a common way to estimate the value function, which can be learned directly from sample data without knowledge of the environment. It is also possible to update the value estimate of the current state by using the value estimates of the subsequent states.

The timing difference update of the Q-Learning algorithm is shown in equation (10):

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{10}$$

After executing the action, the action value function and policy are updated, where α denotes the learning rate and γ denotes the discount factor.

2.3.5. Scheduling Policy Generation. Based on the above process, the process of enterprise multi-objective operation to solve the scheduling strategy is shown in Fig. 3, which includes the parameter optimization of PDR tree and Q-learning. Firstly, the most suitable initial PDR is determined by simulation data, and the PDR tree structure is generated based on the PDR. Based on the tree structure, the parameters, i.e., the initialization strategy, are initialized, and then after L iterations, the intelligent body completes learning. In each iteration round the intelligent body needs to make T actions, T is the number of non-leaf nodes of the PDR tree. Firstly, it needs to obtain the current environment state S. Based on the previously learned knowledge and the current environment, the intelligent body makes action A, gets reward R, and changes the current environment. After completing all T actions, the scheduling target value, i.e., the maximum completion time, is updated, after which the scheduling policy is updated.

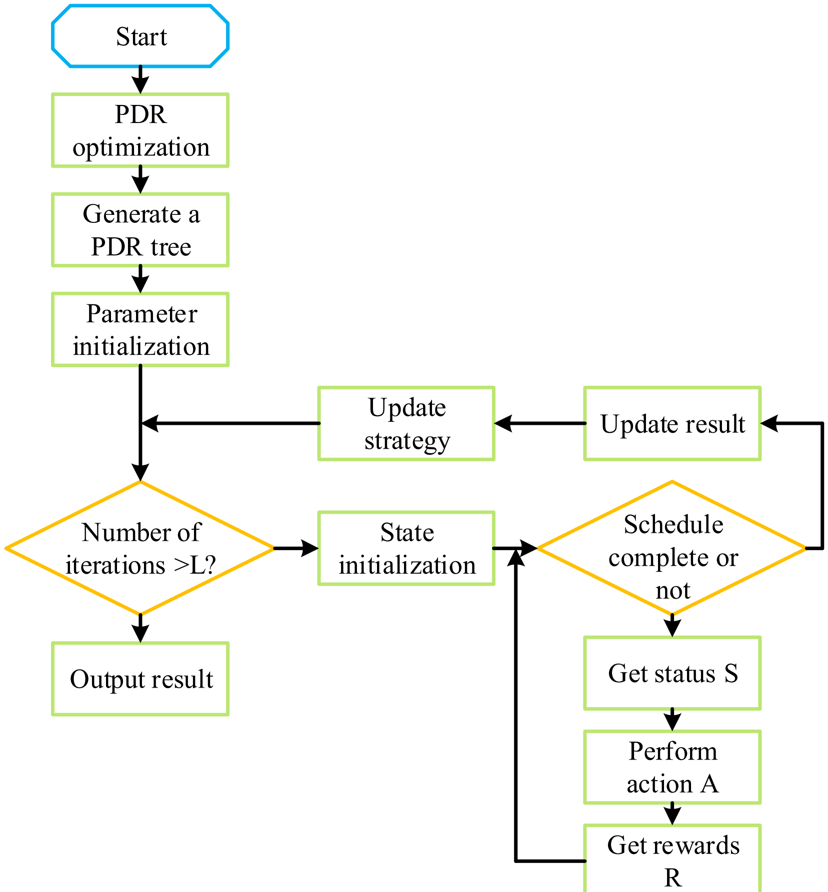

Fig. 3. Algorithm implementation process

2.4. Research Sample and Data Sources

2.4.1. Research sample. The data obtained from listed enterprises are more accurate because listed companies disclose their quarterly and annual financial data every year, and there are also many databases to organize and count their data. According to the Results of Industry Classification of Listed Companies in 2Q2018, and the sample selection of previous scholars' research on the efficiency of listed companies, 50 listed companies are screened as the selection sample. However, this type of classification is rather rough and cannot reflect the real characteristics of listed enterprises. Therefore, this paper excludes ST and *ST listed enterprises as well as enterprises listed and shell-listed after 2010, eliminates enterprises with imperfect data and some missing data, and meets the requirements that each DMU is positive and the proportion of business revenue is not less than 70%. Finally, 23 listed enterprises are identified as the research objects of this paper.

2.4.2. Data sources. This paper takes 2018-2024 as the time period for research and analysis, and the data are obtained from the annual reports disclosed by each listed enterprise and the database of Cathay Pacific, which is denoted as Testdatabase.By comparing the mean and standard deviation of each index in each year, it reflects the effect of the enterprises' multi-objective operation approach based on reinforcement learning algorithms on the improvement of operational efficiency from the side.

3. Results and analysis

3.1. Operational Scheduling Performance Comparison

In order to be able to evaluate the strengths and weaknesses of the scheduling of a company's operational policies from various aspects, the following scheduling policy performance evaluation metrics are used in this paper respectively:

Average: denotes the mean value of the business lag time (in seconds) being dispatched to the enterprise over a period of time.

Std: denotes the variance of the variation of the business lag time (in seconds) being dispatched to the enterprise over a period of time.

Max: the maximum value of the business lag time (in seconds) dispatched to the enterprise over a period of time.

Min: Indicates the minimum value of the time lag (in seconds) of the services dispatched to the enterprise over a period of time.

The smaller the above four metrics are, the better the performance of the company's operation strategy.

Rule-based scheduling strategy (RS), Genetic Algorithm-based scheduling strategy (GES), Decima-based scheduling strategy (DS) are selected with the method of this paper, and the performance of this paper's model and the existing company operation decision-making models are compared on the dataset Testdatabase. Figure 4 shows the scheduling results of different company operation strategies on the dataset Testdatabase. As can be seen from the figure, the company operation strategy models proposed in this paper all exhibit excellent performance in the dataset. On the real dataset, the model in this paper achieves significant improvements in the Average, Std, Max, and Min key performance metrics, which are 123,567, 48,227, 136,830, and 468s, respectively. This is due to the fact that the model training objective focuses mainly on the reduction of the the average operational lag time of the enterprise.

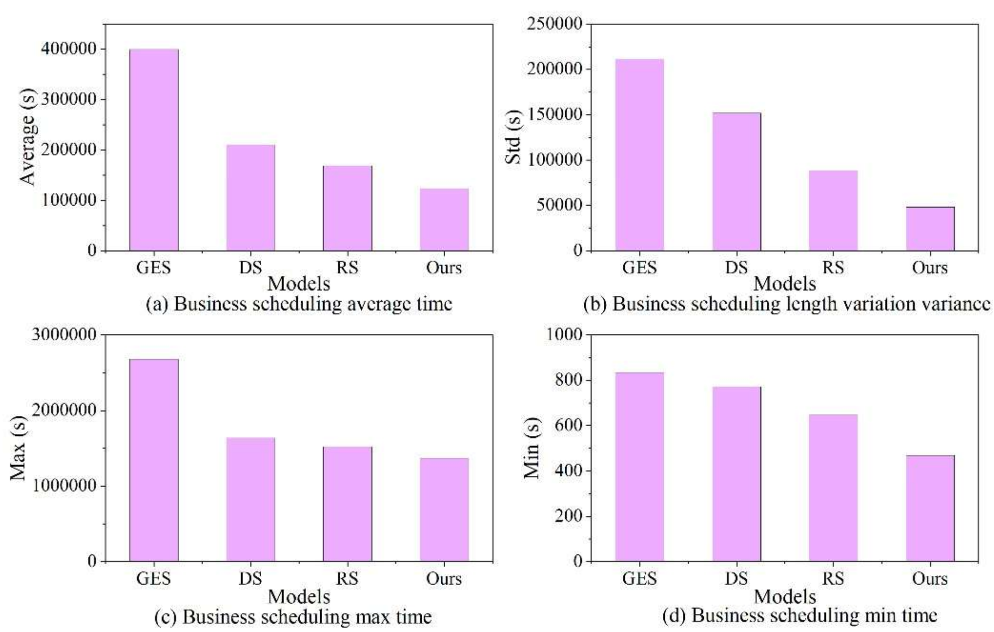

Fig. 4. Scheduling results of different warehouse decisions

In order to further validate the performance robustness of the enterprise operation decision models in this paper, this paper shows the performance comparison results of different business decision models for different scheduling time windows on the Testdatabase dataset. As shown in Fig. 5, it shows the variation of the benefits of different corporate operation decision models for consecutive multiple scheduling time windows on the dataset. Traditional non-reinforcement learning corporate operational strategy models, such as RS and GES, face a common challenge: their business decision benefits fluctuate widely across time windows. The fundamental reason for such fluctuations is that these strategies do not fully consider their long-term impact on the effectiveness of business decisions in subsequent time windows when formulating business decisions in a certain time window, resulting in the fact that decisions in certain time windows may adversely affect the efficiency of business decisions in subsequent time windows. In contrast, the company's operational strategy based on the reinforcement learning algorithm in this paper demonstrates higher stability. This is due to the fact that the model in this paper can predict the demand changes in future time windows based on the current known business demands of corporate operations, and optimize the business decisions in each time window using the reinforcement learning-based approach, with a view to achieving the best business utilization efficiency.

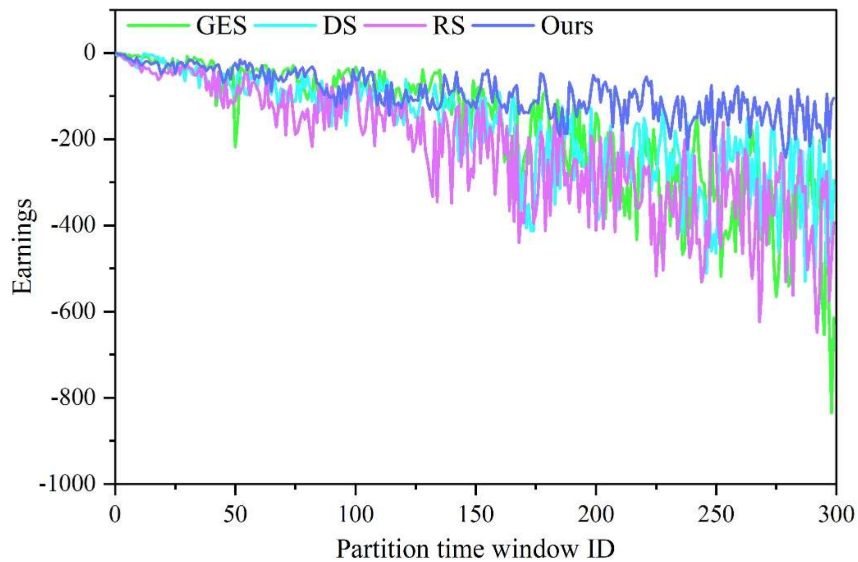

Fig. 5. The income change of each warehouse decision strategy in different time

3.2. *Overall efficiency of business operations*

According to the DEA model, 10 enterprises in a city are taken as samples to implement the operational efficiency improvement strategy designed in this paper. Taking the number of employees, fixed asset investment, and total energy consumption as input indicators, and the main business revenue and freight turnover as output indicators, the operational efficiency of the 10 enterprises from 2018 to 2024 is measured using Excel software, and the operational efficiency in 2025 is predicted. In which the operational efficiency improvement strategy is implemented on January 1, 2021. The operational efficiency of enterprises studied in this paper is the comprehensive efficiency, which can be further split into the corresponding pure technical efficiency as well as scale efficiency. By using the DEA-CCR model, the operational comprehensive efficiency of 10 sample enterprises in a city is analyzed and calculated.

Figure 6 shows the changes in the pure technical efficiency of each enterprise from 2018 to 2025. According to the data in the figure, the pure technical efficiency of the 10 enterprises listed is analyzed, and before the implementation of the strategy of this paper (2018~2020), the average value of the pure technical efficiency of the 10 sample enterprises ranges from 0.3 to 0.4, which indicates that the pure technical efficiency of the 10 enterprises has an average of at least 60% of the upside. However, there are large differences in pure technical efficiency between enterprises, with the mean value varying from 0.209 to 0.583, indicating that the enterprises' pure technical efficiency is not high, their operations are underperforming, and there is more room for improvement overall. After the implementation of the strategy of this paper (2021~present), the pure technical efficiency of each enterprise has improved significantly, and the average value of the pure technical efficiency of the enterprise increases year by year. By 2024 it has reached about 0.9, and it is predicted that it can increase to 0.974 in 2025.This indicates that the efficiency improvement path based on the reinforcement learning algorithm in this paper is able to dynamically adjust the production plan according to the implementation of the production situation and changes in demand to achieve the maximization of production efficiency.

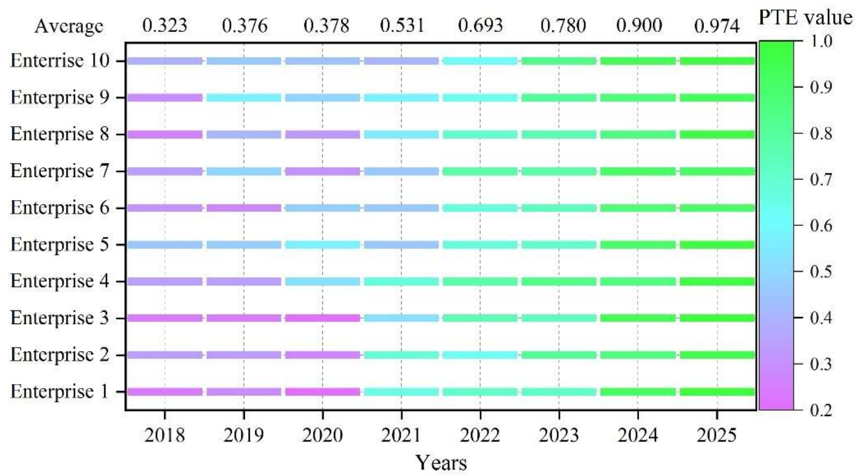

Fig. 6. The pure technical efficiency of the enterprises in 2018~2025

Scale efficiency reflects whether scale size matches resource allocation and whether the optimization between inputs and outputs is achieved. Figure 7 demonstrates the changes in scale efficiency of the ten enterprises from 2018 to 2025. From Figure 7, it can be seen that the change in scale efficiency of each enterprise has a similar trend to the change in pure technical efficiency. Before for the implementation of the strategy of this paper, the scale efficiency of enterprises is between 0.2 and 0.6, which is generally low. After 2021, the scale efficiency shows a clear upward trend, and after 2024, the scale efficiency of each enterprise generally exceeds 0.9.This further demonstrates the effectiveness of this paper's strategy in improving the operational efficiency of enterprises, i.e., using this paper's path to analyze past sales data and market trends to predict future demand changes. Accordingly, inventory levels and replenishment strategies are optimized to improve the overall efficiency and responsiveness of supply scale.

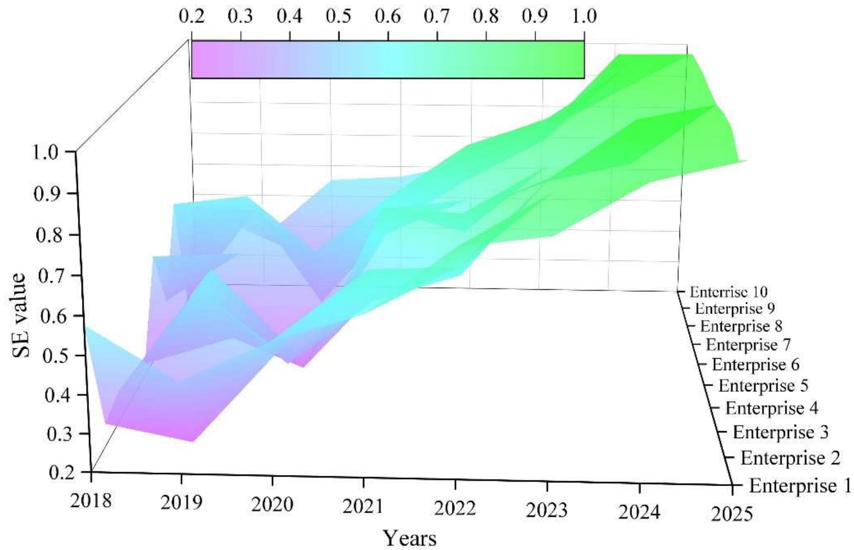

Fig. 7. The change of the scale efficiency of the 10 companies in 2018~2025

3.3. Exploring the relationship between input indicators and operational efficiency

With reference to the indicator rating systems established by many scholars, this section selects reference indicators for the evaluation of operational efficiency of the above 10 companies. In this section, net fixed assets, total employee compensation, operating costs, and administrative expenses are selected as the input indicators of business operations. The specific descriptions of each indicator are as follows:

Net fixed assets: used to represent capital input, fixed capital invested in the process of enterprise operation.

Total Employee Compensation: used to represent the input of labor force.

Operating Costs: used to support the normal operation of the enterprise and to maintain the basic operating facilities of the enterprise.

Administrative costs: used to represent the costs incurred by the enterprise to organize and manage production and operation activities.

By collecting the raw data of 10 enterprises in the above indicators in 2018~2024, the impact of this paper's model on the above indicators is specifically analyzed. Figure 8 shows the average value and relative error of net fixed assets, total employee compensation, operating costs, and administrative expenses of 10 enterprises in 2018~2024. From the figure, it is known that before the implementation of the model of this paper, the operating costs as well as the administrative expenses of the 10 enterprises are high, while the total employee compensation and fixed assets are at a relatively low level. After the adoption of the path of this paper (2021), the average operating costs and administrative expenses of the enterprises are reduced, respectively, 68.29, 19.51. While the employee compensation and fixed assets are significantly higher, respectively, 49.58% and 19.48% higher than in 2020. Combined with the analysis of the results above, it can be seen that operating costs, administrative expenses, employee compensation and fixed assets are closely related to the operational efficiency of the enterprise. Enterprises use reinforcement learning algorithms to consider these factors comprehensively. Multi-objective measures such as optimizing the cost structure, improving management efficiency, motivating employees and improving fixed assets for the enterprise improve the overall operational efficiency of the enterprise.

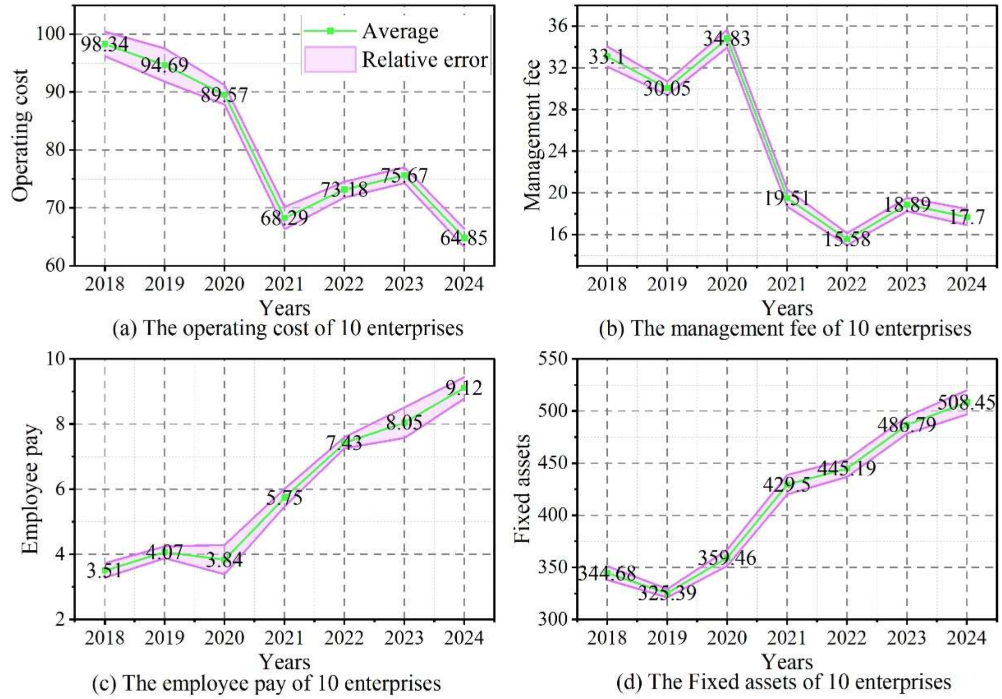

Fig. 8. The average and relative error of 10 companies in 2018~2024

3.4. *Malmquist Index Analysis of Sample Firms*

The previous section was a cross-sectional comparison of the operating efficiency of the 10 firms through the results of the efficiency values. This section introduces the method of Malmquist index and applies it to measure the total factor productivity (TFP) of the 10 firms over a 7-year period to analyze the trend of the whole industry. In addition, the decomposition of total factor productivity is used to analyze in detail the intrinsic reasons for the changes in TFP.

Figure 9 shows the change of Malmquist index of 10 enterprises. The meanings of the indicators presented in the figure are as follows: Effch represents the change in technical efficiency, which represents the change in technical efficiency relative to the previous period, when this indicator is greater than 1, it means that the unit is closer to the frontier surface compared to the previous period. Changes in technical efficiency can be broken down into changes in pure technical efficiency (Pech) and changes in scale efficiency (Sech).Techch represents the rate of technical progress, which is greater than 1, meaning that the production frontier for inputs and outputs has moved to a higher degree. In conjunction with this study, it represents the technical progress of the firm. When Techch is greater than 1, it means that the contribution of technical progress to the productivity index has increased, and vice versa is a decrease in the contribution.The source of the Malmquist index is the movement of technical efficiency as mentioned above, the movement of technical progress and scale efficiency.

As a whole, the TFP index of the 10 firms is greater than 1 over the 7-year period, which means that their operational efficiency has improved over the 7-year period. In terms of the change of technical efficiency, 8 out of 10 enterprises have a change of more than 1, accounting for 80%, which indicates that the distance between most of the enterprises and the frontier side of the industry is shrinking, and most of the enterprises improve their input-output efficiency under the intervention of the model in this paper. And on the indicator of technical progress, all enterprises have shown technical progress in 7 years, which indicates that the path proposed in this paper promotes the production frontier surface of the whole industry in a good direction. Finally, from the change of scale efficiency, as many as 9 out of 10 enterprises have a change of more than 1, which indicates that in the industry, reinforcement learning has helped most of the enterprises in the state of incremental scale payoffs, and there is still considerable room for the development of the whole industry.

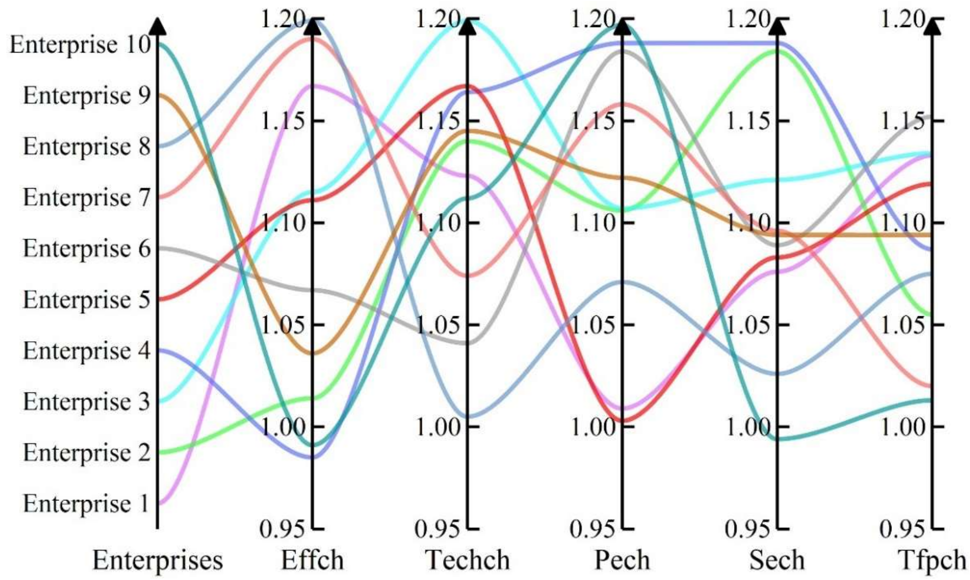

Fig. 9. Changes in the Malmquist index of 10 enterprises

4. Conclusion

In this paper, we propose a Q-learning scheduling-based path for improving the efficiency of enterprise operations. The path generates a PDR tree structure, which turns the enterprise operations into an ensemble representation, and solves it by an algorithm. The Q-learning algorithm is utilized to learn directly in the enterprise operation data by making actions, getting rewards, changing the environment, and deriving the optimal enterprise operation scheduling strategy. The changes in the efficiency of enterprise operations under reinforcement learning are analyzed in conjunction with the enterprise operations measurement method proposed in this paper.

1) The method in this paper significantly reduces the business lag time of the enterprise with Average, Std, Max, and Min of 123,567, 48,227, 1368,310, and 468s, respectively.

2) Firms applying the strategy of this paper increase their pure technical efficiency as well as scale efficiency year by year, and in 2024, the efficiency value of several firms exceeds 0.9.

3) The reinforcement learning algorithm optimizes the cost structure and management system of the enterprises, and improves the overall operational efficiency of the enterprises.

4) The Malmquist indexes of all 10 enterprises under the path of this paper have been improved, and the TFP indexes of all of them are greater than 1, which means that the operational efficiency has been significantly improved.

Currently, reinforcement learning algorithms require a large amount of interaction data to train the model for optimal decision making. Obtaining such a large amount of high-quality data is a huge challenge in business operations. And reinforcement learning optimizes decisions through constant trial and error, which can cause irreversible losses in some areas. This requires algorithm improvement to increase the convergence speed and sample utilization of the algorithm through improved learning mechanisms and optimized algorithm structure. To make its application in business operations more extensive and efficient.

References

[1] Szepesvári, C. (2022). Algorithms for reinforcement learning. Springer nature.

[2] Shakya, A. K., Pillai, G., & Chakrabarty, S. (2023). Reinforcement learning algorithms: A brief survey. Expert Systems with Applications, 231, 120495.

[3] Oh, J., Hessel, M., Czarnecki, W. M., Xu, Z., van Hasselt, H. P., Singh, S., & Silver, D. (2020). Discovering reinforcement learning algorithms. Advances in Neural Information Processing Systems, 33, 1060-1070.

[4] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[5] Gumerov, A., Ismagilov, R., Akhmedzyanova, F., & Akhmetov, M. (2020). Increasing labour productivity at industrial enterprises. European Proceedings of Social and Behavioural Sciences.

[6] Radionova, N., Skrypnyk, M., & Voronkova, T. (2019). Dual nature of industrial enterprise cost management system. Baltic Journal of Economic Studies, 5(2), 184-190.

[7] Park, K. J., & Yoo, Y. (2017). Improvement of competitiveness in small and medium-sized enterprises. Journal of Applied Business Research, 33(1), 173.

[8] Shvachych, G., & Kholod, E. (2017). Research of the enterprise profitability strategy. Baltic Journal of Economic Studies, 3(5), 451-456.

[9] Babafemi, I. D. (2015). Corporate strategy, planning and performance evaluation: A survey of literature. Journal of Management Policies and Practices, 3(1), 43-49.

[10] Baumgartner, R. J., & Rauter, R. (2017). Strategic perspectives of corporate sustainability management to develop a sustainable organization. Journal of Cleaner Production, 140, 81-92.

[11] Wadström, P. (2019). Aligning corporate and business strategy: managing the balance. Journal of Business Strategy, 40(4), 44-52.

[12] Djurayeva, M. S. (2024). ORGANIZATIONAL STRUCTURE OF A MODERN ENTERPRISE. Ethiopian International Journal of Multidisciplinary Research, 11(12), 222-227.

[13] Zusman, P. (2021). Participants' Ethical Attitudes and Organizational Structure and Performance: Application to the Cooperative Enterprise. In Agricultural cooperatives in transition (pp. 23-54). Routledge.

[14] Bellini, C. G. P., Pereira, R. D. C. D. F., & Becker, J. L. (2016). Organizational structure and enterprise systems implementation: Theoretical measures and a benchmark for customer teams. Information Technology & People, 29(3), 527-555.

[15] Dolynskyi, S. V., Kostina, T. Y., & Voskolupov, V. V. (2024). THE ESSENCE OF THE ENTERPRISE MANAGEMENT SYSTEM AND FEATURES OF ITS STRUCTURE. Electronic scientific and practical publication in economic sciences.

[16] Vuta, D. R., & Farcas, A. (2015). The role of training in organizational and employee development. Land Forces Academy Review, 20(3), 367.

[17] Chunhui, C., & Jantan, A. H. B. (2023). Research on the Relationship between Enterprise Human Resource Training and Employee Job Performance. Journal of International Business and Management, 6(8), 01-12.

[18] Asgarova, A. (2019). Impact of employee training on employee job satisfaction and achievement: a case of Turkish manufacturing industry. Available at SSRN 3398117.

[19] Lin, T. H., & Xiong, Y. (2017). The Existing Problems and Countermeasures in the Training of Enterprise Staff. Destech Transactions on Computer Science and Engineering.

[20] Löhr, K., Weinhardt, M., Graef, F., & Sieber, S. (2017). Enhancing communication and collaboration in collaborative projects through conflict prevention and management systems. Organizational Dynamics, 47(4), 259-264.

[21] Banaeianjahromi, N., & Smolander, K. (2019). Lack of communication and collaboration in enterprise architecture development. Information Systems Frontiers, 21, 877-908.

[22] Schiller, C., & Meiren, T. (2018, June). Enterprise social networks for internal communication and

collaboration. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1-4). IEEE.

[23] Wang, F. J., Chiu, W., Tseng, K. F., & Cho, H. (2023). The roles of employee–employee collaboration and employee–customer collaboration in fitness service innovation: a comparison of frontline and non-frontline employees. International Journal of Sports Marketing and Sponsorship, 24(4), 793-813.

[24] Schmeleva, A. N. (2017). Evaluation and improvement of the operating efficiency of enterprise quality management system (QMS): Conceptual bases. Calitatea, 18(160), 100.

[25] LUBNINA, A. A., Shinkevich, M. V., Yalunina, E. N., Gaidamashko, I. V., Savderova, A. F., & Komissarova, M. A. (2018). Innovative strategy for improving the efficiency of industrial enterprises management. Revista Espacios, 39(09).

[26] Jiao, R., Zhao, G., & Wang, F. (2019). LOGISTICS SERVICE STANDARDIZATION, ENTERPRISE OPERATION EFFICIENCY, AND ECONOMIC EFFECTS. Transformations in Business & Economics, 18(3).

[27] Gao, L., Kim, G. J., & Kim, H. H. (2019). A Study on the Evaluation of Operational Efficiency of E-Commerce Demonstration Enterprise in China. Journal of Digital Convergence, 17(12), 181-190.

[28] Qiao, L., & Fei, J. (2022). Government subsidies, enterprise operating efficiency, and “stiff but deathless” zombie firms. Economic Modelling, 107, 105728.

[29] Madanhire, I., & Mbohwa, C. (2016). Enterprise resource planning (ERP) in improving operational efficiency: Case study. Procedia CIRP, 40, 225-229.

[30] Lin, B., & Xie, Y. (2023). Does digital transformation improve the operational efficiency of Chinese power enterprises?. Utilities Policy, 82, 101542.

[31] Li, T., Wen, J., Zeng, D., & Liu, K. (2022). Has enterprise digital transformation improved the efficiency of enterprise technological innovation? A case study on Chinese listed companies. Math. Biosci. Eng, 19(12), 12632-12654.

[32] ZHOU, L., LI, W., TEO, B. S. X., & YUSOFF, S. K. M. (2022). The effect of green transformation on the operating efficiency of green M&A enterprises: Evidence from China. The Journal of Asian Finance, Economics and Business, 9(1), 299-310.

[33] Xin Yao & Jiajia Zhao. (2024) .The impact of big fund on the operational efficiency of semiconductor companies: Corporate social responsibility as a non-economic output. International Journal of Production Economics109429-109429.

[34] S. Allaei,Behnam Vahdani,H. R. Gholami & A. Alinezhad. (2024). Assessing cross-docking performance using a novel network DEA model regarding sustainability and undesirable factors under uncertainty. International Journal of Systems Science: Operations & Logistics(1).

[35] Mahya Jodeiri Rad & Costas Armenakis. (2024). Active Reinforcement Learning for the Semantic Segmentation of Urban Images. Canadian Journal of Remote Sensing(1).

[36] Lakshmi Mandal,Chandrashekar Lakshminarayanan & Shalabh Bhatnagar. (2025). Approximate linear programming for decentralized policy iteration in cooperative multi-agent Markov decision processes. Systems & Control Letters106003-106003.

[37] Shuyun Zhang,Huifang Chen & Lei Xie. (2025). Service-aware Q-learning-based routing protocol in the underwater acoustic sensor network. Computer Networks110986-110986.