PLANT SCIENCE TODAY
ISSN 2348-1900 (online)
Vol 11(sp4): 01–08
https://doi.org/10.14719/pst.5426

HORIZON
e-Publishing Group
HePG

RESEARCH ARTICLE

Adaptive routing in agricultural supply chains: Harnessing Q-learning for optimal decision-making in dynamic environments

Mantaw Suliyaa Chow[1], M Prahadeeswaran[1*], V Karthick[1], CS Sumathi[2] & Patil SG[2]

[1] Department of Agricultural Economics, Tamil Nadu Agricultural University, Coimbatore 641 003, India
[2] Department of Physical Sciences and Information Technology, Agricultural Engineering College, Tamil Nadu Agricultural University, Coimbatore 641 003, India

*Email: prahadeeswaranmecon@tnau.ac.in

OPEN ACCESS

ARTICLE HISTORY

Received: 01 October 2024
Accepted: 15 October 2024

Available online
Version 1.0 : 10 December 2024

Check for updates

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

CITE THIS ARTICLE

Chow MS, Prahadeeswaran M, Karthick V, Sumathi CS, Patil SG . Adaptive routing in agricultural supply chains: Harnessing Q-learning for optimal decision-making in dynamic environments. Plant Science Today.2024;11(sp4):01-08.
https://doi.org/10.14719/pst.5426

Abstract

In this study, the authors try to emphasize how Q-learning, a model-free reinforcement learning (RL) technique can be used for optimizing routing in a grid-based environment. This study aims to assess the efficacy of Q-learning in enhancing routing for agricultural supply chains, investigate its flexibility in dynamic environments, and compare its performance across several real-world scenarios. In this specific case of the banana chain, an agent is moving through various entities in the system - from local growers to small traders and warehouses. It models the routing problem as a Markov Decision Process (MDP) and the goal is to optimize cumulative reward. Several possible cases are simulated, e.g. the finding of an optimal route for a given visit sequence that optimizes charging time and non-drivable paths left over when unexpected blockages occur to avoid energy wear penalties as well as how to best save costs; These results demonstrate the adaptability and durability of Q-learning in dynamic environments to obtain near-optimal solutions across diverse settings. Indeed, the present study adds to a growing body of research on the application of RL in logistics and supply chain management, highlighting its potential to enhance decision-making in complex and variable environments. The findings suggest that Q-learning can effectively balance multiple objectives, such as minimizing distance, reducing costs, and avoiding high-wear areas, making it a valuable tool for optimizing routing in real-world supply chains. Future work will explore broader applications and other RL algorithms in similar contexts.

Keywords

Markov Decision Process (MDP); logistics; Q-learning; routing optimization

Introduction

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have unlocked new opportunities for optimizing complex decision-making processes in dynamic environments. A significant area of interest is RL, particularly within logistics and supply chain management. RL, a subset of ML, provides a framework for an agent to learn optimal policies through interactions with an environment. Among the various RL algorithms, Q-learning is notable for its model-free nature, enabling the agent to learn directly from raw experiences without requiring a model of the environment. Q-learning has been widely applied to routing problems, where the objective is to identify the most efficient path in each environment. These

problems are often modelled as a Markov Decision Process (MDP), which represents the environment through a set of states, actions, transition probabilities, and rewards (1). In this context, the goal of Q-learning is to discover an optimal policy that maximizes the expected cumulative reward, thereby enabling the agent to make decisions that lead to the most favorable outcomes over time (1).

Traditional routing algorithms, such as Dijkstra's and the A* algorithm, require a complete and static representation of the environment to function effectively. However, real-world scenarios are frequently characterized by uncertainty and dynamic changes, such as road blockages, fluctuating traffic conditions, and varying transportation costs (2). These challenges necessitate the use of adaptive algorithms capable of responding to changes in real-time. Q-learning, with its ability to learn from continuous interactions with the environment, is particularly well-suited for such applications (3). This research focuses on applying Q-learning to routing within a grid-based environment, specifically modeled on the banana supply chain. The environment is represented as a grid, where each cell corresponds to a different stakeholder in the supply chain, such as farmers, local traders, and warehouses. The agent's task is to navigate from a starting point to a goal, passing through various intermediate points while optimizing multiple objectives, including minimizing distance, reducing costs, and accounting for wear and tear (4). In the grid-based model, each cell corresponds to a specific stakeholder or location within the banana supply chain. For instance, cells might represent local farms, trading centers, storage warehouses, or distribution hubs, reflecting the real-world journey of bananas from harvest to consumer. This gives a resemblance to the real-world representation of the banana supply chain.

Q-learning is a widely used RL algorithm due to its simplicity and model-free nature, allowing it to learn optimal policies without requiring a model of the environment (5). Unlike SARSA, which is an on-policy algorithm that updates its action-value function based on the actual actions taken, Q-learning is off-policy, updating based on the maximum reward across all possible actions, leading to more exploration and potentially better long-term policies (6). In contrast, Deep Q-Networks (DQN) utilize deep neural networks to approximate Q-values, enabling them to handle larger and more complex state spaces (7). However, DQNs come with higher computational demands and require careful tuning to ensure stability, which can be challenging in dynamic environments. Q-learning's computational efficiency and straightforward implementation make it particularly suitable for problems with well-defined, discrete states-such as grid-based agricultural supply chains where rapid adaptation to changes is critical.

The study examines multiple situations to assess the efficacy of the Q-learning paradigm. These situations comprise optimal route selection based on distance, reactions to unforeseen route obstructions, assessment of penalties arising from substandard road conditions, and tactics for cost reduction by circumventing high-cost routes (8). By simulating these scenarios, the research aims to demonstrate the flexibility and robustness of Q-learning in solving real-world routing problems. Previous studies have demonstrated the potential of RL in dynamic environments, such as warehouse management and robotic path planning (9). However, there remains a need for more research into the practical deployment of these algorithms in logistics and supply chains, particularly in sectors like agriculture, where conditions are highly variable. This work contributes to this growing body of research by applying Q-learning to a complex routing problem within the agricultural supply chain, providing insights into how RL can enhance operational efficiency and decision-making in dynamic environments.

Materials and Methods

Traditional MDP assume known transition probabilities. In contrast, Q-learning learns these probabilities through interactions with the environment, making it suitable for scenarios with unknown explicit probabilities and values. The MDP framework, introduced by A.A. Markov (1856-1922), is widely used in dynamic systems modeling (10). In the context of the banana supply chain, states are defined as the different locations within the supply chain grid, such as farms, warehouses, or trading posts. Actions correspond to the possible movements between these locations, and rewards are calculated based on factors like transportation costs, distance traveled, and penalties for delays, thus tailoring the MDP to the dynamics of agricultural logistics. The key components of the MDP framework include State (S), Actions (A), and Transition versus future rewards. Probabilities (P), Reward (R), and Discount Factor (γ)

The optimal reward is given by the Bellman equation given as

$$V(s) = max_a(R(s,a) + \gamma V(s') \quad \text{.....(Eqn-1)}$$

V(s) : Expected return value at the current state's', max_a: The maximum value of any possible action 'a', *R(s, a)*: The expected reward for taking action 'a' at state 's', *yV(s')*: The discount factor gamma multiplied by the value of the next state.

The choice of γ plays in determining an optimal reward, in this study, the discount factor (γ) was set to [0.9] a standard value. A higher γ emphasizes long-term rewards, encouraging the agent to plan routes that are optimal over time, while a lower γ focuses on immediate rewards, favoring quick and possibly suboptimal paths. The chosen value strikes a balance between these factors, aiming to optimize routes that account for both short-term efficiency and long-term benefits, γ lies between 0 to 1 (inclusive). If γ is set to 0 the V(s') term is negated completely and the model only cares about the immediate reward, if γ is set to 1, the model weights potential future rewards as equal to the immediate rewards.

A Simulated case scenario of common stakeholders involved in the banana supply chains with their routes and their possible connections was made (Fig 1). A grid-based approach was used to address a RL routing problem. The agent navigates from the base (B) to the wholesaler (WS) through various stakeholders i.e. Farmer (F), Local Trader (LT), Primary Processing center (PPC), Central Warehouse (CW), and Distributor (D), actions include directional movements on a 4x4 grid (Fig 2).

Bellman equation that enables the agent to learn the optimal policy though it does not need a model of the environment. This approach is especially used in applications where the state-action space is discrete and the agent aims to learn a policy that maximizes reward over time.

Optimal Path Finding

The problem of optimal pathfinding within a grid environment can be modeled as a MDP, where the environment is

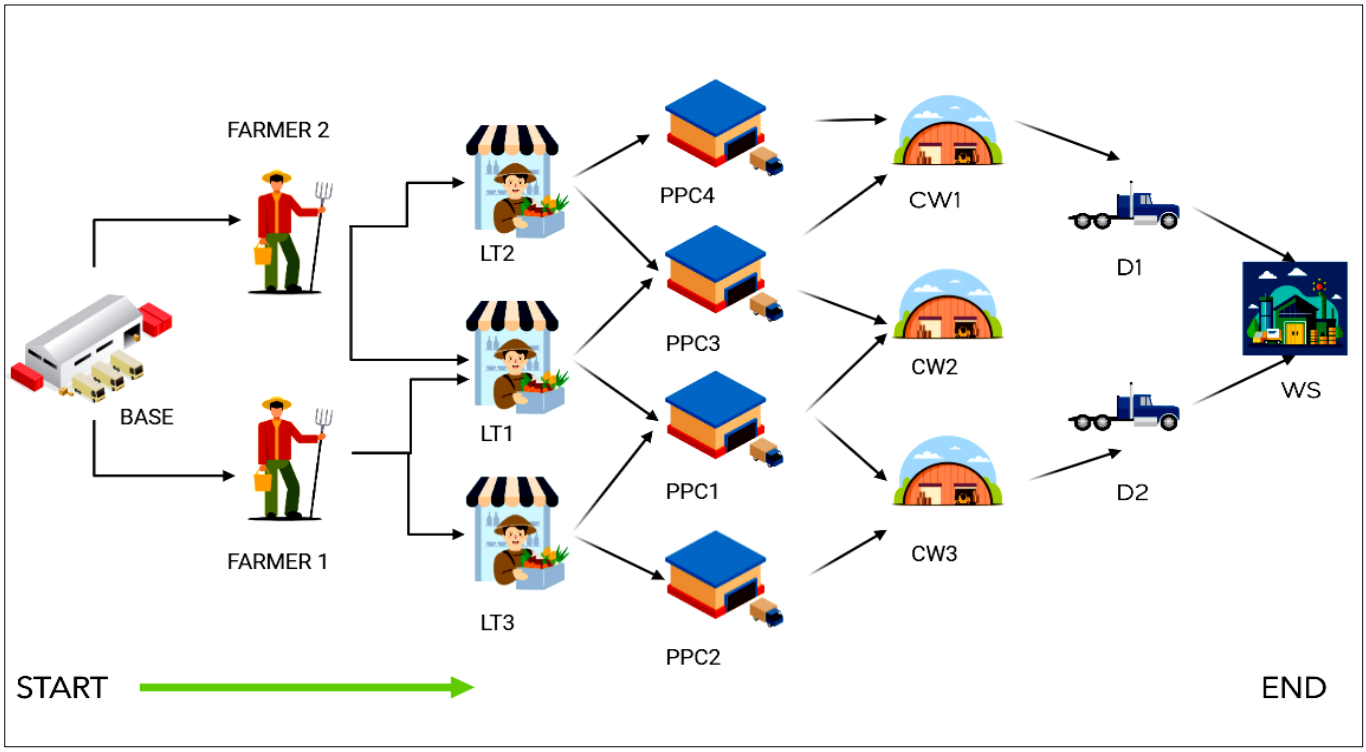

Fig. 1. The Network diagram of all the possible routes for the Problem.

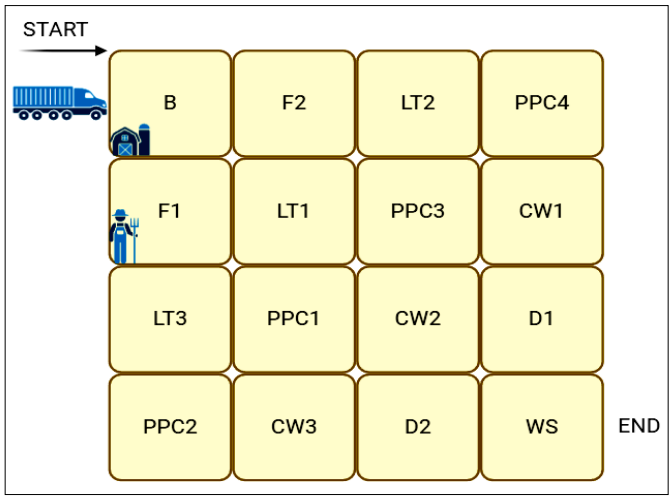

Fig. 2. A 4x4 grid-based depiction of the stakeholders and their route connections.

The problem is solved using a RL problem with Q-learning using Open AI GYM Space in a python environment (11). Furthermore, Q-learning is a model-free RL algorithm that helps an agent to learn the best policy for making the decisions in each environment. It does so by using a Q-value function that is then iteratively updated which determines the sum of the expected future reward when a given action is taken in a particular state. In performing these updates, the Q-learning algorithm uses the

defined by a tuple (S, A, P, R, γ).

The state space S is defined as the set of all possible positions on the grid where n is the size of the grid and each state corresponds to a specific S_{ij} cell in the grid, defined as:

$$S = \{S_{ij} \mid i, j \in \{1,2,\dots,n\}\} \quad \text{.....(Eqn-2)}$$

Action space A consists of four possible movements, up, down, left and right, where a_1, a_2, a_3, a_4 represents the direction of movement respectively, these actions can be represented as

$$A = \{a_1, a_2, a_3, a_4\} \quad \text{.....(Eqn-3)}$$

The transition probability P function defines the probability of transitioning from one state S_t to the next state, S_{t+1} given an action a_t . For deterministic environments,

$$P(s_{t+1} \mid s_t, a_t) = \begin{cases} 1 & if\ s_{t+1} = deterministic\ outcome\ of\ s_t\ after\ taking\ a_t \\ 0 & otherwise \end{cases}$$

.....(Eqn-4)

P is defined as

Reward function $R(S_t, a_t)$ assigns a scalar value based on the action taken in a given state, where r is the

$$R(s_t, a_t) = \begin{cases} r \ \ if\ a_t\ leads\ to\ a\ state\ closer\ to\ the\ goal \\ -r \ \ if\ a_t\ leads\ to\ a\ state\ farther\ from\ the\ goal \end{cases}$$

.....(Eqn-5)

magnitude of the reward. It can be expressed as:

The policy $\pi(a \mid s)$ defines the strategy used by the agent to select actions. The optimal policyπ * is the one that maximizes the expected cumulative reward, here$Y$ is

$$\pi^*(s) = arg\ max_a \sum_{S'} P\,(s'|s,a)[R(s,a) + \ \gamma.V^*(s')]$$

.....(Eqn-6)

the discount factor, and $V^*(s')$ is the value of the next state. It is defined as:

The Q-learning algorithm that is used to iteratively update the action-value function $Q(s,a)$ which estimates the expected utility of taking action a in state s . Here α is the learning rate, and Y is the discount factor.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\,[R(s_t, a_t) + \gamma\, max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)]$$

.....(Eqn-7)

The equation is defined as:

After training, the optimal policy π * is used to derive the optimal path from the initial state S_0 to the goal state S_g ,

$$Path = \{s_0, s_1, \ldots, s_g\}\ where\ s_{t+1} = \pi^*(s_t)$$

.....(Eqn-8)

the path is visualized by marking the paths visited according to π *, this step is defined as:

Scenarios

Scenario 1

Optimal route selection based on distance, involving all stakeholders.

Scenario 2

Response to sudden route blockages leading to CW3.

Scenario 3

Consideration of wear and tear penalties set with constant “k” on certain routes due to poor road conditions. The weightage for wear and tear was set at 0.9 and for distance at 0.2.

Scenario 4

Cost reduction by avoiding high-cost routes due to tolls and wear and tear.

Each scenario reflects common challenges in agricultural supply chains: sudden blockages simulate unpredictable disruptions like road closures or accidents; wear and tear penalties represent varying road conditions affecting vehicle maintenance; and cost considerations reflect the need to balance transportation efficiency with economic factors, such as toll costs and fuel prices.

Results

This section provides the detailed results of the Q-learning Problem, The Optimal routes along with their optimal results, and Total rewards help us understand that the model generalizes well, Results in Table 1 and 2 are the optimized and worst-case results respectively. The Initial grid in Fig 2. Shows the starting problem, the grid is the same for all scenarios, Different parameters in Different Scenarios will determine, the optimal route for the agent.

Scenario 1

Different Distance Involved: The RL model has determined an optimal route through the given grid, as indicated by the highlighted cells. The best route identified by the model starts at B, then proceeds through F1, LT3, PPC1, CW3, and D2, and finally ends at WS. This Highlighted path (Fig 3 a) represents the most efficient trajectory through the various locations or waypoints in the grid. The model achieved this solution with optimal parameters, resulting in a Total Reward of -60 and a total distance of 168 km for the entire route, with the worst case being taking a route that would lead to a longer distance of 208 km.

Scenario 2

A blockage in Waypoint CW3: The RL model has identified an alternative optimal route through the grid, accounting for a blockage at CW3. The new best route, highlighted in the image, begins at B, then proceeds through F1, LT3, PPC1, CW2, D1, and finally reaches WS. This path (Fig 3 b) represents the most efficient trajectory given the new constraint of the blocked CW3 location. Despite the obstruction, the model maintained the same Total Reward of -60, although the total distance increased to 185 km. This outcome demonstrates the model's ability to adapt to changes in the environment while still optimizing for the defined reward function. The slight increase in total distance compared to the previous scenario (from 168 km to 185 km) reflects the necessary detour to avoid the blocked area while still achieving an efficient route through the required waypoints. Here, in worst-case scenario if the agent follows the worst path, it follows a similar route to scenario 1, thus leading to a longer distance of 208 km.

Scenario 3

Wear and Tear Penalty: The wear and tear penalties are designed to mimic the impact of poor road quality, challenging terrain, and vehicle depreciation over time. These factors are critical in agricultural logistics, where maintaining vehicle conditions can significantly affect operating costs and delivery reliability. The RL model has determined an optimal route through the grid while considering wear and tear penalties. The highlighted path starts at B, then proceeds through F2, LT1, PPC1, CW3, and D2, and finally ends at WS. This route (Fig 3 c) avoids the areas with very high wear and tear, which are depicted by the cross-

Table 1. Optimized results of the MDP problem in each scenario.

Scenario	Total Reward	Optimized result	Optimal Route						
1	-60	168 km (Total distance)	B	F1	LT3	PPC1	CW3	D2	WS
2	-60	185 km (Total distance)	B	F1	LT3	PPC1	CW2	D1	WS
3	-51.60	186 km (Total Distance); k = 24 wear and tear penalty	B	F2	LT1	PPC1	CW3	D2	WS
4	-60	173 km (Total Distance) ₹4610 Total cost	B	F2	LT1	PPC1	CW2	D1	WS

Table 2. Results show the maximum possible worst-case for each scenario.

Scenario	Worst case result	Worst Route						
1	208 km (Total distance)	B	F2	LT2	PPC4	CW1	D1	WS
2	208 km (Total distance)	B	F2	LT2	PPC4	CW1	D1	WS
3	165 km (Total Distance); k = 105 wear and tear penalty	B	F1	LT3	PPC1	CW2	D2	WS
4	200 km (Total Distance); ₹6490 total cost	B	F1	LT1	PPC3	CW1	D1	WS

shaded cells (F1, LT3, and CW2). The model achieved an improved optimal reward of -ve 51.60, compared to previous scenarios, with a total distance of 186 km. The total wear and tear constant k = 24, which factors into the overall optimization. This outcome demonstrates the model's ability to balance multiple objectives - minimizing distance, maximizing reward, and now also considering the impact of wear and tear on different path segments. The slight increase in total distance compared to the initial scenario is offset by the improved reward, likely due to the avoidance of high-wear areas, showcasing a more nuanced approach to path optimization in this complex environment. However, in the worst case, if the agent takes the worst path, then the total distance is reduced to 165 km, although hugely increasing the wear and tear penalty from k = 24 to k = 105.

Scenario 4

A Cost-based approach: The cost-based analysis incorporates factors such as toll charges, fuel expenses, and potential penalties for delays. These components are crucial for logistics companies seeking to minimize expenses while maintaining timely deliveries, particularly in agricul-

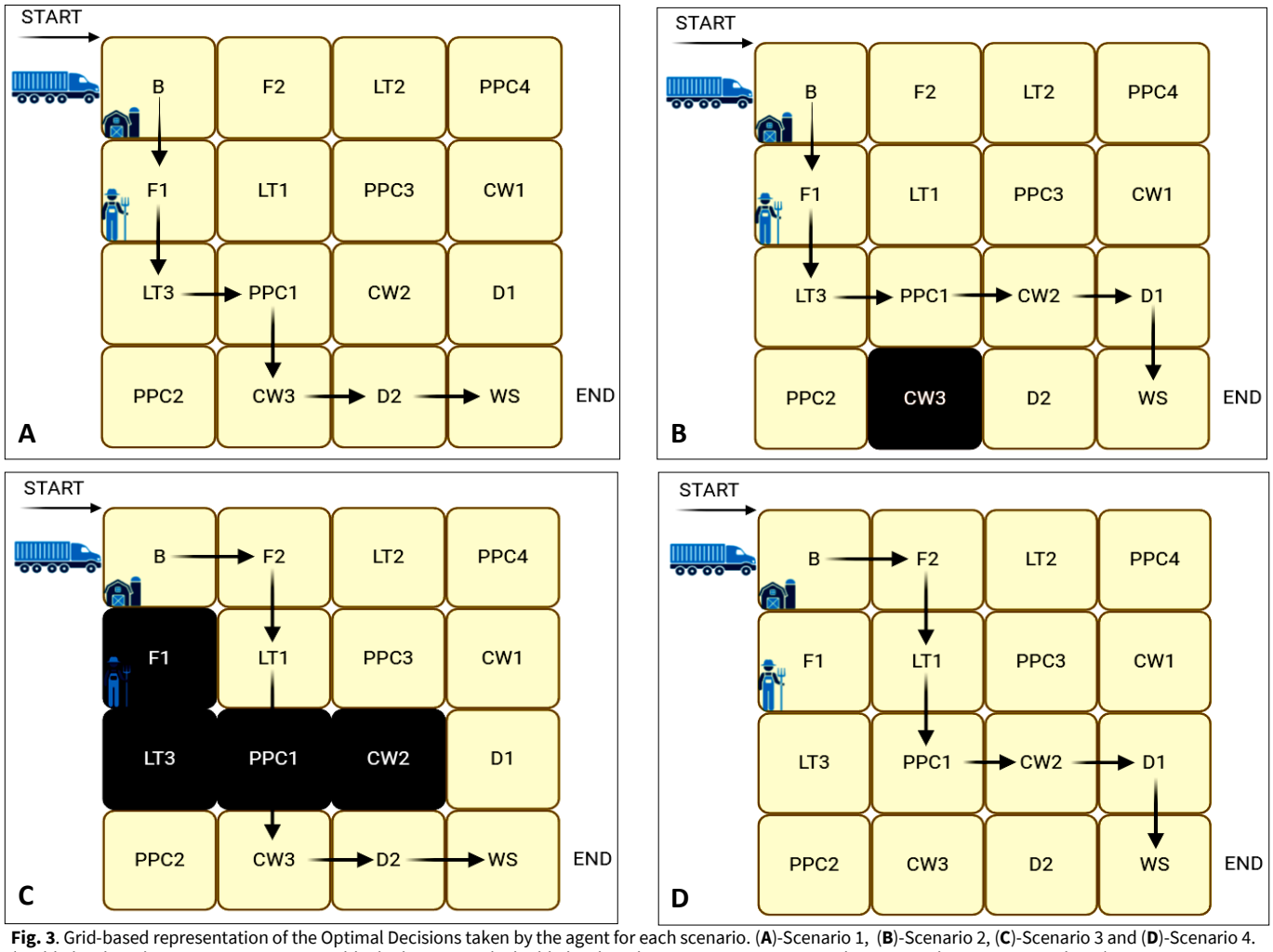

Fig. 3. Grid-based representation of the Optimal Decisions taken by the agent for each scenario. (**A**)-Scenario 1, (**B**)-Scenario 2, (**C**)-Scenario 3 and (**D**)-Scenario 4. (Highlighted marks in scenario 2 represent blockade in CW3; The highlighted marks in scenario 3 represent the areas with more wear and tear).

tural contexts where margins are often narrow. In this new scenario, the RL model has identified an optimal route through the grid based on cost considerations. The highlighted path begins at B, continues through F2, LT1, PPC1, CW2, and D1, and finally reaches WS. This route (Fig 3 d) represents the most cost-effective trajectory through the various locations or waypoints in the grid. The model achieved a solution with a total cost of ₹4610 rupees and a Total Reward of -60. This outcome suggests that the RL model has balanced minimizing costs with maximizing the reward function. The highlighted cells indicate the chosen path, which likely represents a trade-off between the shortest distance and the least expensive route. The model has successfully navigated through the grid while optimizing for both cost and reward, demonstrating its ability to adapt to different optimization criteria in complex routing problems. Furthermore, in worst-case scenario, the agent will accumulate a distance of 200 km and an increase in cost from ₹4610 to ₹6490.

This outcome suggests that compared to classical routing algorithms like Dijkstra's, which require a static and complete map of the environment, Q-learning demonstrates greater flexibility. Its ability to adapt to real-time changes, such as sudden blockages, allows it to outperform static methods, especially in dynamic environments. Unlike Dijkstra's algorithm, Q-learning continually updates its policy based on new information, optimizing routes even under unpredictable conditions. The RL model has balanced minimizing costs with maximizing the reward function. The highlighted cells indicate the chosen path, which likely represents a trade-off between the shortest distance and the least expensive route. The model has successfully navigated through the grid while optimizing for both cost and reward, demonstrating its ability to adapt to different optimization criteria in complex routing problems.

This is true because the form of rewards demonstrated in Fig 4 is episodic confirms that the RL model minimizes penalties in all the analyzed scenarios and shows the corresponding penalties' decrease. This has shown the generality of the model in being able to handle several constraints and optimization criteria.

Limitations

This study's main limitation is its reliance on a discrete, grid-based environment, which may not capture the full complexity of real-world supply chains with continuous variables. Additionally, the model assumes a simplified reward structure that might not fully represent economic trade-offs in a more complex logistics network. Future work could involve integrating continuous state spaces and incorporating a wider range of economic and environmental variables to enhance the model's robustness.

Discussion

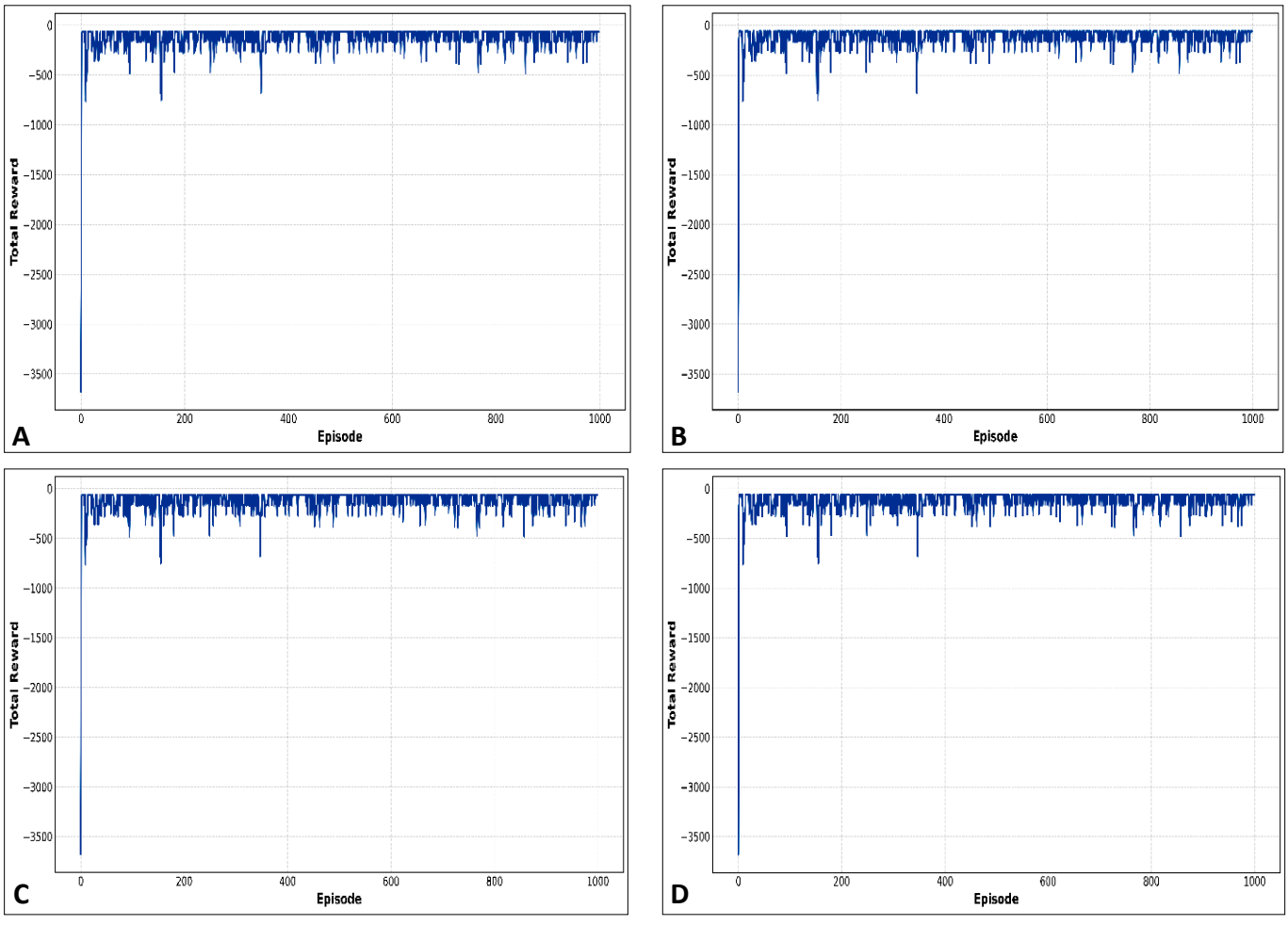

Fig. 4. Episodic rewards of the different scenarios used, which shows, the model running well. (**A**)-Scenario 1, (**B**)-Scenario 2, (**C**)-Scenario 3 and (**D**)-Scenario 4.

The RL model's effectiveness in route allocation was evaluated through various scenarios, each with unique constraints and objectives. The RL model identified an optimal route starting at B, proceeding through F1, LT3, PPC1, CW3, D2, and ending at WS. The total reward achieved was -60, with a total distance of 168 km. This scenario highlights the model's ability to find the most efficient path in a grid-based environment, optimizing for distance and reward. With CW3 blocked, the model rerouted through B, F1, LT3, PPC1, CW2, D1, and WS, maintaining the same reward of -60 but increasing the distance to 185 km, the model demonstrated adaptability to environmental changes, effectively rerouting to maintain efficiency despite obstacles. Considering wear and tear penalties, the model identified a route through B, F2, LT1, PPC1, CW3, D2, and WS, achieving an improved reward of -51.60 and a total distance of 186 km, with a wear and tear constant of 24. This scenario illustrates the model's capability to balance multiple objectives, including minimizing distance, maximizing reward, and avoiding high-wear areas.

The RL model optimized for cost, selecting a path through B, F2, LT1, PPC1, CW2, D1, and WS, resulting in a total cost of ₹4610 and a reward of -60. This scenario highlights the model's proficiency in optimizing routes based on cost considerations, achieving a balance between cost efficiency, and reward. The RL model's ability to adapt to various scenarios, such as blockages and wear-and-tear considerations, is supported by research on RL in dynamic environments.

The RL model consistently adapted to changing environmental conditions, such as blockages and wear and tear zones, finding optimal or near-optimal solutions. It successfully balanced multiple objectives, including distance, reward, wear and tear, and cost, showcasing its potential for real-world applications. Given that the RL model is flexible and resistant to change, it will fit well into the logistic, transportation, and route planning scenarios where conditions and priorities change constantly. The use of RL can be effectively applied in logistics to optimize decisions throughout the supply chain and gain various advantages. Discussion on the effectiveness of RL in multi-objective function problems like distance minimization, reward maximization, and cost reduction is evident. A study on deep RL for logistics task coordination showed that RL algorithms could generalize solutions for various tasks, balancing efficiency and cost considerations effectively (12). The model's ability to adapt to various scenarios and balance multiple objectives underscores its value in complex, dynamic environments.

APPENDIX

Dataset

In this study, the application of RL, with a focus on Q-learning, demonstrated its effectiveness in addressing route optimization problems within grid environments. The RL models were capable of adapting to dynamic conditions, including variations in route parameters and the introduction of constraints such as blockages and wear-and-tear penalties. This adaptability underscores the potential of RL to enhance operational efficiency and cost-effectiveness, particularly in complex logistical scenarios. The success of Q-learning in this context suggests that RL can be a powerful tool for optimizing logistics and supply chain management, especially in sectors where dynamic conditions frequently impact operations, such as agriculture. Further research should explore other RL algorithms and techniques to push the boundaries of optimization in logistics and supply chains. Additionally, expanding the scope to include international market dynamics and global supply chain issues could provide valuable insights, offering a more comprehensive perspective on the challenges and opportunities in supply chain management on a global scale. This could lead to the development of more robust and universally applicable models, further enhancing the relevance and impact of RL in various industrial and agricultural contexts.

Acknowledgements

A very heartful thanks to Virag G, and Vishwajeet Avhale for helping in preparing this manuscript, a special gratitude towards TNAU, Coimbatore, and the Department of Agricultural Economics Coimbatore. CARDS to provide computing facilities to complete this research with ease.

Authors' contributions

MSC conceived the study, carried out data curation, formal analysis, and investigation, developed the methodology, and drafted the original manuscript. MP supervised the project, contributed to the investigation, and participated in validation and original draft writing. VK participated in conceptualization and contributed to the review and editing of the manuscript. CSS contributed to the investiga-

From	To	Scenario 1 (Distance in Km)	Scenario 2 (Distance with Blockade in CW3)	Scenario 3 (Wear and tear penalty)	Scenario 4 (Cost in ₹)
B	**F1**	30	30	20	1100
B	**F2**	50	50	2	780
F1	**LT1**	35	35	4	1090
F1	**LT3**	20	20	20	800

F2	**LT1**	20	20	2	620
F2	**LT2**	25	25	3	780
LT1	**PPC1**	28	28	5	890
LT1	**PPC3**	32	32	6	1020
LT2	**PPC3**	28	28	5	890
LT2	**PPC4**	40	40	7	1270
LT3	**PPC1**	30	30	20	1100
LT3	**PPC2**	35	35	7	1120
PPC1	**CW2**	40	40	20	1210
PPC1	**CW3**	38	X	7	1410
PPC2	**CW3**	20	X	3	630
PPC3	**CW1**	55	55	11	1760
PPC3	**CW2**	30	30	5	950
PPC4	**CW1**	45	45	9	1440
CW1	**D1**	28	28	4	880
CW2	**D1**	15	15	2	470
CW2	**D2**	20	X	20	800
CW3	**D2**	25	X	3	780
D1	**WS**	20	20	4	640
D2	**WS**	25	25	5	800

tion, participated in validation, and reviewed and edited the manuscript. PSG participated in the conceptualization, investigation, validation, visualizations, and review and editing of the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest: The Authors of This paper declare no conflict of Interest.

Ethical issues: None

References

1. Watkins CJCH, Dayan P. Q-Learning. Mach Learn.1992;8:279-292 https://doi.org/10.1007/BF00992698
2. Pallottino S, Scutellà MG. Shortest path algorithms in transportation models: classical and innovative aspects. In: Marcotte P, Nguyen S, editors. Equilibrium and Advanced Transportation Modelling: Centre for Research on Transportation. Springer, Boston, MA; 1998.p245-81. https://doi.org/10.1007/978-1-4615-5757-9_11.
3. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529-33. https://doi.org/10.1038/nature14236
4. Tijms HC. A first course in Stochastic models. John Wiley & Sons, Ltd;2003.
5. Sewak M. Temporal difference learning, SARSA and Q-learning. Deep Reinforcement Learning.Springer;2019. https://doi.org/10.1007/978-981-13-8285-7
6. Azar NA, Shahmansoorian A, Davoudi M. Uncertainty-aware path planning using reinforcement learning and deep learning methods. Journal of Computer and Knowledge Engineering.2020;3(1):25-37.
7. Rodrigues P, Vieira SM. Optimizing agent training with deep Q-learning on a self-driving reinforcement learning environment. 2020 IEEE Symposium Series on Computational Intelligence (SSCI). 2020:745-52. https://doi.org/10.1109/SSCI47803.2020.9308525
8. Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd ed. A Bradford Book, Cambridge; 2018. (edited based on the link https://www.scirp.org/reference/referencespapers?referenceid=2465216)
9. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Driessche GVD, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484-9. https://doi.org/10.1038/nature16961
10. White III CC, White DJ. Markov decision processes. European Journal of Operational Research. 1989 Mar 6;39(1):1-6.
11. Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W. Openai gym. arXiv 2016. arXiv preprint arXiv:1606.01540. 2020.
12. Chenatti S, Previato G, Cano G, Prudencio R, Leite G, Oliveira T, et al. Deep reinforcement learning in robotics logistic task coordination. In 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR), and 2018 Workshop on Robotics in Education (WRE) 2018; 326-332. https://doi.org/10.1109/LARS/SBR/WRE.2018.00066