

Article

Using Reinforcement Learning in a Dynamic Team Orienteering Problem with Electric Batteries

Majsa Ammouriova ^{1,2}, Antoni Guerrero ^{3,4}, Veronika Tservadze ³, Christin Schumacher ⁵ and Angel A. Juan ^{3,6,*}

¹ Industrial Engineering Department, German Jordanian University, Amman 11180, Jordan; maysa.ammouri@gju.edu.jo

² Computer Science Department, Universitat Oberta de Catalunya, 156 Rambla Poblenou, 08018 Barcelona, Spain

³ Research Center on Production Management and Engineering CIGIP, Universitat Politècnica de València, Plz. Ferrandiz-Salvador, 03801 Alcoy, Spain; antoni.guerrero@baobabsoluciones.es (A.G.); vtservadze@upv.es (V.T.)

⁴ Baobab Soluciones, 55 Jose Abascal, 28003 Madrid, Spain

⁵ The Department of Business and Economics, TU Dortmund University, 44221 Dortmund, Germany; christin.schumacher@tu-dortmund.de

⁶ Department of Business Analytics, Euncet Business School, 1 Camí Mas Rubial, 08225 Terrassa, Spain

* Correspondence: ajuanc@upv.es

Abstract: This paper addresses the team orienteering problem (TOP) with vehicles equipped with electric batteries under dynamic travel conditions influenced by weather and traffic, which impact travel times between nodes and hence might have a critical effect on the battery capacity to cover the planned route. The study incorporates a novel approach for solving the dynamic TOP, comparing two solution methodologies: a merging heuristic and a reinforcement learning (RL) algorithm. The heuristic combines routes using calculated savings and a biased-randomized strategy, while the RL model leverages a transformer-based encoder-decoder architecture to sequentially construct solutions. We perform computational experiments on 50 problem instances, each subjected to 200 dynamic conditions, for a total of 10,000 problems solved. The results demonstrate that while the deterministic heuristic provides an upper bound for rewards, the RL model consistently yields robust solutions with lower variability under dynamic conditions. However, the dynamic heuristic, with a 20 s time limit for solving each instance, outperformed the RL model by 3.35% on average. The study highlights the trade-offs between solution quality, computational resources, and time when dealing with dynamic environments in the TOP.

Keywords: team orienteering problem; battery management; electric vehicle; reinforcement learning



Citation: Ammouriova, M.; Guerrero, A.; Tservadze, V.; Schumacher, C.; Juan, A.A. Using Reinforcement Learning in a Dynamic Team Orienteering Problem with Electric Batteries. *Batteries* **2024**, *10*, 411. <https://doi.org/10.3390/batteries10120411>

Academic Editor: Karim Zaghib

Received: 30 September 2024

Revised: 9 November 2024

Accepted: 22 November 2024

Published: 25 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the rise of electric vehicles (EVs) and unmanned aerial vehicles (UAVs), vehicle routing challenges have gained new relevance in modern disaster, security, and last-mile logistic planning [1,2]. The growing popularity of EVs has initiated a transformative shift in the automotive, vehicle, and logistics industry, providing a sustainable alternative to traditional internal combustion engine vehicles. EVs have the potential to reduce transportation costs and minimize environmental impact. In logistics, EVs must navigate complex delivery routes while managing their limited battery life and the necessity for frequent recharging, which introduces an additional layer of complexity to routing, visiting points, and delivering goods. Including battery constraints and charging cycles in modeling is essential for the practical deployment of these vehicles, whether they are delivering goods in urban areas, conducting aerial surveillance, or responding to emergencies [3]. For instance, in urban logistics, the challenge is optimizing the delivery schedule and ensuring that vehicles can complete their routes without interruptions. This problem becomes even more complex when multiple vehicles have different energy requirements [4].

In these scenarios, it is critical to optimize the coverage of regions within a limited time frame [5]. The primary challenge lies in determining optimal paths that prioritize the most relevant areas while considering the energy constraints of the vehicles. This becomes particularly complex when the power infrastructure is compromised and charging opportunities are limited [6]. These challenges can be modeled as team orienteering problems (TOPs), where the EV fleet needs to deliver goods to specific points in the area, starting at the depot. The TOP is a well-established combinatorial optimization problem, extending the classical orienteering problem [7] by introducing multiple agents or vehicles. Each team member attempts to visit candidate nodes within a prescribed time limit. Visiting a node for the first time allows the collection of a reward, and the goal is to maximize the overall team score. The problem is NP-hard even in its single-agent form, and it becomes even more difficult to solve when multiple agents are involved [8]. In real-life scenarios, these vehicles face several challenges due to dynamic conditions, such as those associated with weather status, traffic status, and travel times [9].

In scenarios where real-time routing plans are required, the ability to continuously re-optimize routes as new data become available is key. For example, if a traffic accident occurs or extreme weather conditions arise, the planned routes for EVs or UAVs may need to be adjusted on the fly. The TOP, combined with agile optimization techniques, provides a robust framework for handling these dynamic conditions, ensuring that the vehicles can continue to operate efficiently despite the changing environment [10]. By taking advantage of the power of parallel computing and biased randomization techniques, agile optimization algorithms can concisely explore a vast solution space, identifying routes that meet immediate operational needs and optimize energy usage and overall efficiency. This paper explores a dynamic TOP where vehicles must travel from an origin to a destination to maximize the number of nodes visited within the constraints of battery life. The objective is to optimize the route planning such that vehicles can collect as much reward as possible before their batteries are depleted. Figure 1 shows a dynamic scenario where an EV must travel from an origin to a destination, visiting as many nodes as possible within battery life constraints. The green circles represent visited nodes, while the pink circles are non-visited nodes. The travel time between nodes is influenced by dynamic factors such as weather conditions and road congestion, indicated by 'Tdynamic' in the figure.

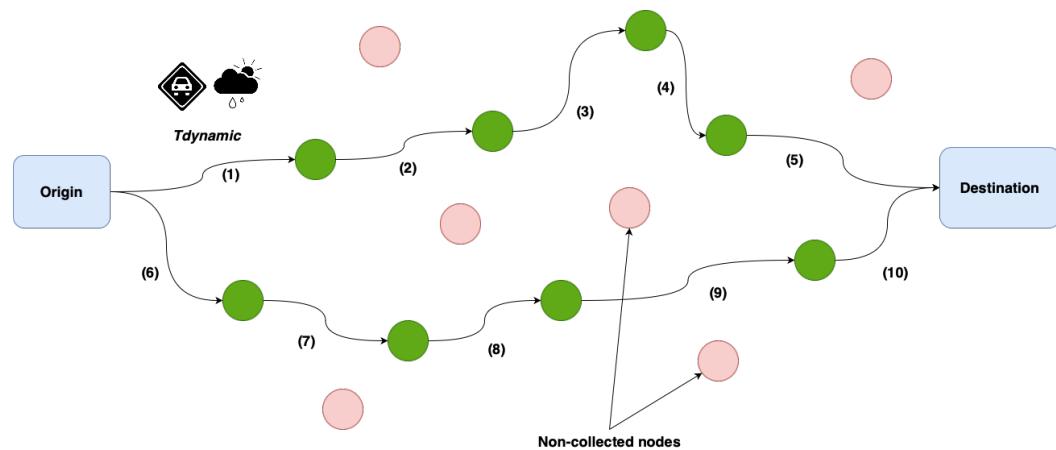


Figure 1. TOP with multiple vehicles considering battery constraints and variable travel times.

To address this dynamic version of the TOP with electric batteries, this paper proposes two approaches: one based on a biased-randomized heuristic [11] and the other using reinforcement learning (RL) [12]. The selection of reinforcement learning is primarily due to its possible suitability for dynamic problems like the TOP, where travel times fluctuate. RL allows the solving agent to adaptively learn and optimize routes under these changing conditions, potentially improving solution adaptability and quality over deterministic methods. Moreover, one of the paper's main objectives was to evaluate the

performance of RL in comparison to more established approaches. The heuristic provides a solution by optimizing routes based on predefined rules, while the neural network in the RL approach is trained separately to learn and refine decisions in response to dynamic conditions. Both methods are independently tested and compared in terms of their ability to adapt to changes such as varying battery levels and travel conditions. Hence, the main contribution of this paper is the development and comparison of two approaches, heuristic-based and RL-based, for solving a TOP involving EVs with dynamic rewards. This dual approach allows us to evaluate the strengths and weaknesses of each method in making adaptive decisions in a rapidly changing environment. The remainder of this paper is structured as follows: Section 2 offers an analysis of the different EV categories, their integration into existing transportation systems, current advancements in the field, as well as factors influencing battery performance, including capacity, materials, and charging times. Section 3 summarizes the related work on algorithms designed for solving TOPs. Section 4 formally defines the specific problem. Section 5 presents the dataset used for the study. Section 6 outlines the simulation experiments and the RL algorithm applied. Section 7 details and interprets our computational findings. Finally, Section 8 concludes with a summary of the main findings and recommendations for future research.

2. Sustainable Logistics and EVs

Sustainable transportation has been included in the United Nations Sustainable Development Goal 11 (<https://sdgs.un.org/topics/sustainable-transport>, accessed on 29 September 2024). The transportation sector influences society development through its economic, social, and environmental impacts [13]. Additionally, it contributes heavily to gas emissions and fuel consumption. According to data collected by the European Commission (<https://ec.europa.eu/eurostat/web/main/data/database>, accessed on 29 September 2024), the transportation sector consumed around 31% of the total energy consumed in 2022 in the European countries, while the industry sector consumed around 25% in the same year. The transportation sector released 23% of energy-related emissions in 2019 [14]. In recent decades, sustainable transportation has become a topic for researchers investigating logistics and chain management, new fuels and vehicles, public transportation improvements, and sustainable transport alternatives [13]. These research efforts aim to optimize energy consumption in transportation and reduce gas emissions. As a result, a variety of concepts have been defined, such as ride-sharing and car-sharing, and new transportation options have been introduced to the market. EVs form one of these options aiming to reduce gas emissions as well as noise in cities [15]. In these vehicles, the conventional internal combustion engine is replaced by another one utilizing electricity from a rechargeable battery to drive the vehicle [16]. EVs are pictured as innovative symbols for decarbonizing transportation with new business opportunities and models [17]. The introduction of EVs improves air quality in terms of reducing greenhouse gas emissions and decreasing operation costs [16]. Hence, a sustainable urban environment is being developed in cities.

Several EVs types are found in the market: hybrid electric vehicles (HEVs), plug-in hybrid electric vehicles (PHEVs), battery electric vehicles (BEVs), and fuel cell electric vehicles (FCEVs). Multiple energy generation and conversion technologies are utilized in HEVs [18]. HEVs use an internal combustion engine or an electrical motor to drive the vehicle depending on the driving condition regulated by a control system. The coupling configuration of the internal combustion engine and the electric motor could be series, parallel, or series-parallel, as illustrated by Cao et al. [19]. Batteries in these vehicles are utilized to store electricity used to drive the electric motors. Real-time optimization methods are utilized to control the energy flow, including charging and discharging of the batteries. The greatest challenge in this type of EVs is the management of multiple energy sources [18]. Another type of hybrid electric vehicle is the PHEV. Similar to HEVs, PHEVs combine internal combustion engine and an electrical motor. Batteries in PHEVs might not only be charged during the drive, but they can be charged directly from the grid

network [20]. BEVs utilize only electricity to drive the vehicle motor (electrical motor) [21], and the battery increases both vehicle weight and price. The driving range of BEVs depends on their batteries, which forms one of the main challenges compared to other EVs [18]. In contrast to BEVs, FCEVs have powerful input and long driving range [22]. FCEVs are EVs fueled by hydrogen that is used to generate the electricity [23]. Thus, FCEVs are supplied with hydrogen fuel tank that forms one of the challenges associated with this vehicle type (safety concerns). According to Pramuanjaroenkij and Kakaç [23], fuel cell technology is the future of transportation, with many ongoing studies.

Batteries in EVs vary in their technology, resulting in different charging speeds, driving range, and safety issues. The first kind of batteries are lead–acid batteries [16,18], consisting of a lead electrode and electrolyte. They have low gravimetric energy density (30–50 Wh/kg) and can be used for small vehicles [18]. Later, nickel-based batteries were invented [16,18], with a higher gravimetric energy density, e.g., 60–80 Wh/kg for a nickel–cadmium battery and around 140 Wh/kg for a nickel–hydrogen battery. In addition, this type of battery has a long life cycle and can be recharged many times. Many HEVs are supplied with nickel-based batteries. Lithium-ion batteries were invented in the second half of the last century [18] and are used in energy storage systems. They are characterized by a high gravimetric energy density (118–250 Wh/kg) and longer battery life. Different lithium-ion batteries types are found in the market [24]. Additionally, these batteries types varies in their cost and charging time. Various safety design guidelines for high-energy-density batteries have been explored [25]. Advances in battery chemistry, recycling, charging infrastructure, and cost reduction are continuous. The research concerning battery chemistry reflects the invention of various battery types. Since some types of EVs are being charged with electricity, charging infrastructure have been studied [16]. The location of charging outlets has become an issue for drivers using EVs. Related to this issue, recharging speed is another ambiguous aspect. The charging speed depends on the battery type, e.g., lithium-ion batteries are recharged relatively quickly. Another approach is replacing batteries in station instead of recharging them [26]. This approach demands that the issues presented by battery interchangeability, brand compatibility, and battery ownership be resolved, which will not be an easy task. Despite these challenges, the exchange of batteries provides a solution for long charging times and the need to upgrade household installations to adapt to fast charging. EVs are characterized by their driving range, which depends upon driving behavior, battery type [18], as well as the surrounding environment defined by the weather conditions, infrastructure, and traffic [27,28]. In this context, the state of charge indicates the maximum driving range according to the charge level of the batteries.

The use of EVs raises significant issues related to power distribution grid, environment, and safety. In the pursuit of sustainable development, optimization of battery operation and safety were investigated [29]. The charging of batteries in PHEVs and BEVs causes increased power demand and voltage drops [30]. Innovative approaches target supporting the power distribution grid, such as vehicle-to-grid technology. In these approaches, EVs discharge power saved in their batteries to the grid [31]. Concerning the environment, batteries consist of toxic materials and require proper disposal procedures. Additionally, thermal runaways are a concern that might lead to fires or explosions. Heating and cooling techniques for batteries were also studied [24]. Thermal management systems for batteries are crucial for controlling heating and cooling performance and the stability of batteries.

3. Related Work on Team Orienteering Problems

Traditional approaches to solve TOPs have primarily focused on deterministic and stochastic versions that assume that the environment is either entirely predictable or subject to random variations. However, real-world applications often involve dynamic changes, such as fluctuating weather conditions, varying traffic congestion, and evolving battery status, particularly relevant in the context of EVs [32]. These factors complicate the decision-making process, making it necessary to incorporate real-time data and adap-

tive algorithms into the solution methodology [33]. Recent research has explored various methodologies to optimize vehicle routing in dynamic and stochastic environments. For instance, recent studies have demonstrated the effectiveness of RL in navigating dynamic orienteering challenges, showing that RL-based approaches can outperform traditional heuristic methods in uncertain scenarios [34,35]. To address the complexity of dynamic TOPs, Panadero et al. [36] have investigated the use of RL combined with simheuristics to tackle dynamic and stochastic TOP scenarios, where variables such as battery life and travel times are unpredictable. Other studies have explored the use of deep reinforcement learning (DRL) for multi-vehicle TOP scenarios, demonstrating how DRL can adapt to real-time changes in operational environments [37]. Additionally, some studies have applied RL in a multi-stage TOP, emphasizing scenarios where both rewards and constraints evolve over time and offering insights into long-term planning and decision-making under uncertainty [38].

The dynamic TOP becomes particularly challenging when applied to EVs, which face constraints such as battery life and charging station availability. Effective routing in this context must take these dynamic factors into account to optimize both the travel route and the battery management system (BMS). Studies have highlighted the importance of advanced BMS to improve EV performance and longevity [39,40]. Further research has demonstrated the application of RL in this area, showing how RL can optimize load node selection and route planning under dynamic conditions [41]. Despite significant advances, challenges remain in fully addressing the dynamic TOP using RL. The high computational complexity of these problems, together with the difficulty of obtaining high-quality solutions in short computing times—especially under dynamic and stochastic conditions—continues to drive research toward more efficient algorithms [42].

4. Modeling the TOP with Dynamic Travel Times

The TOP is a classical NP-hard problem where a set of nodes and vehicles are given, and the main goal is to maximize the sum of the reward collected from visiting each node. A maximum travel time is allowed for each vehicle, which means that not all nodes can be visited. This travel time can be limited by various factors, such as the battery capacity of EVs. Moreover, each vehicle has to exit from the initial depot and end its route at the final one. In a formal way, let $G = (N, E)$ be a directed graph, where $N = \{1, 2, \dots, n\} \cup \{o, d\}$ represents the set of nodes, with o and d denoting the origin and destination depots, respectively. The set of directed edges is given by $E = \{(i, j) \mid i, j \in N\}$, which represents the possible paths between nodes. Each node $i \in N$ has an associated reward r_i for visiting that node, with the rewards at the origin and destination depots, r_o and r_d , set to zero. Let V be the set of all vehicles. A binary decision variable, x_{ijv} , takes the value of 1 if vehicle v travels from node i to node j and 0 otherwise. Then, the mathematical model for TOP with dynamic travel times is formulated as follows:

$$\max \sum_{v \in V} \sum_{i, j \in N} x_{ijv} r_j \quad (1)$$

Equation (1) defines the objective function, which aims to maximize the total reward collected from the visited nodes. The following constraints apply:

$$\sum_{j \in N} x_{ojv} \leq 1 \quad \forall v \in V \quad (2)$$

$$\sum_{i \in N} x_{idv} = \sum_{j \in N} x_{ojv} \quad \forall v \in V \quad (3)$$

$$x_{ijv} \leq \sum_{j \in N} x_{ojv} \quad \forall v \in V \quad (4)$$

$$\sum_{i \in N} x_{iov} + \sum_{j \in N} x_{dsv} = 0 \quad \forall v \in V \quad (5)$$

Equation (2) ensures that a vehicle departs the origin at most once. Equations (2) and (3) state that if a vehicle departs from the origin depot, it must eventually arrive at the destination depot. Meanwhile, Equation (4) ensures that a vehicle can visit other nodes only if it leaves the origin depot, and Equation (5) enforces that no vehicle can revisit the origin depot or depart from the destination depot, ensuring that the route starts at the origin and ends at the destination. Additional constraints are described next:

$$\sum_{v \in V} \sum_{i \in N} x_{ijv} \leq 1 \quad \forall j \in N \setminus \{d\} \quad (6)$$

$$\sum_{i \in N} x_{ijv} = \sum_{i \in N} x_{jiv} \quad \forall j \in N \setminus \{o, d\}, v \in V \quad (7)$$

$$y_{iv} - y_{jv} + 1 \leq (1 - x_{ijv})|N| \quad \forall i, j \in N, v \in V \quad (8)$$

Equation (6) ensures that each node is visited at most once, while Equation (7) states that if a vehicle arrives at a node, it must also depart from that node. Moreover, in Equation (8) the variables y_{iv} are introduced, which represent the order of the node i in the route of vehicle v . This restriction ensures that there are no subtours. A last set of equations is introduced next:

$$\sum_{i, j \in N} x_{ijv} f(i, j) \leq L \quad \forall v \in V \quad (9)$$

$$y_{iv} \geq 0, \quad \forall i \in N, v \in V \quad (10)$$

$$x_{ijv} \in \{0, 1\} \quad \forall i, j \in N, v \in V \quad (11)$$

Equation (9) ensures that the travel time for each vehicle does not exceed the maximum allowable travel time L . Here, $f(i, j)$ represents the travel time between nodes i and j , which is dynamic and varies according to the specific conditions of the edge. Equations (10) and (11) enforce the nature of the variables.

5. A Numerical Case Study

Consider a TOP with dynamic travel times, which are influenced by weather and traffic conditions. The positions of the nodes are known, as are the deterministic travel times. Since this is a case study, we are supposing that these travel times can increase by up to 12.5%, depending on the dynamic conditions between each pair of nodes. For illustrative purposes, let us assume that the travel time can be modeled using a linear regression. In practical applications, actual travel data could indeed be leveraged to provide more precise predictions. However, implementing such a model was beyond the scope of this study, which focused on evaluating the two different methods: heuristics and RL models. The weather and traffic values range from 0 to 1, depending on their influence on the travel time increase. Hence, we will assume that the ‘real’ travel time between nodes i and j is given by $f(i, j)$:

$$f(i, j) = t_{ij}(1 + w_{ij} \cdot 0.0625 + h_{ij} \cdot 0.0625)$$

In this case, t_{ij} represents the deterministic travel time between nodes i and j , w_{ij} represents the weather conditions, and h_{ij} represents the traffic conditions between the pair of nodes. A value of 0 indicates favorable weather or traffic conditions, while 1 represents the worst conditions. It is assumed that weather and traffic conditions contribute equally to the travel time increase (each with the coefficient 0.0625). Each node has a reward that also varies between 0 and 1, and the x and y coordinates of each node are randomly generated within the range [0, 1]. Additionally, the maximum travel time allowed for each vehicle is randomly generated, but the minimum allowed travel time is set to be 1.125 times the travel time between the two depots. This function, f , is used solely for calculating travel times and can be more complex to better reflect real-world conditions. However, the exact nature of this function remains unknown to the solving algorithm. In this study, 50 different problem instances have been defined. These problem instances differ in node location

(x and y coordinates), vehicle maximum travel time, and node rewards. A total of 20 nodes are found in each instance, and these nodes can be visited by one of two vehicles to collect rewards from them. Likewise, the dynamic conditions (weather and traffic) have been varied 200 times for each instance. It is important to recall that each problem instance is defined by a specific set of nodes, rewards, and depot locations, and then it is solved 200 times, each one under different dynamic conditions (which means different travel times between node pairs due to changing factors). The reported reward for each instance represents the average reward collected across these 200 different dynamic condition.

6. Solving Approaches

The heuristic approach is inspired by the method proposed by Panadero et al. [8]. It is a merging heuristic, where an initial dummy solution is created, consisting of simple routes from the origin depot to a node and then to the destination depot, and routes are then combined based on calculated savings. The savings $s_{i,j}$ for each pair of nodes i and j are computed as follows:

$$s_{i,j} = \alpha(c_{0,j} + c_{i,n+1} - c_{i,j}) + (1 - \alpha)(r_i + r_j) \quad (12)$$

In this equation, $c_{0,j}$ represents the travel cost from the initial depot to node j , $c_{i,n+1}$ represents the cost of traveling from node i to the destination depot, and $c_{i,j}$ represents the travel cost between nodes i and j . Additionally, r_i and r_j denote the rewards associated with nodes i and j , respectively. The merging process incorporates a biased-randomized strategy [11], meaning that merges are not performed in a purely greedy way, but rather in a randomized manner that aligns with the heuristic's logic. Specifically, a geometric distribution is applied for the biased-randomization, with a parameter β that varies randomly between 0.1 and 0.3. This heuristic, in conjunction with the methodology shown in Figure 2, is used to solve the dynamic TOP.

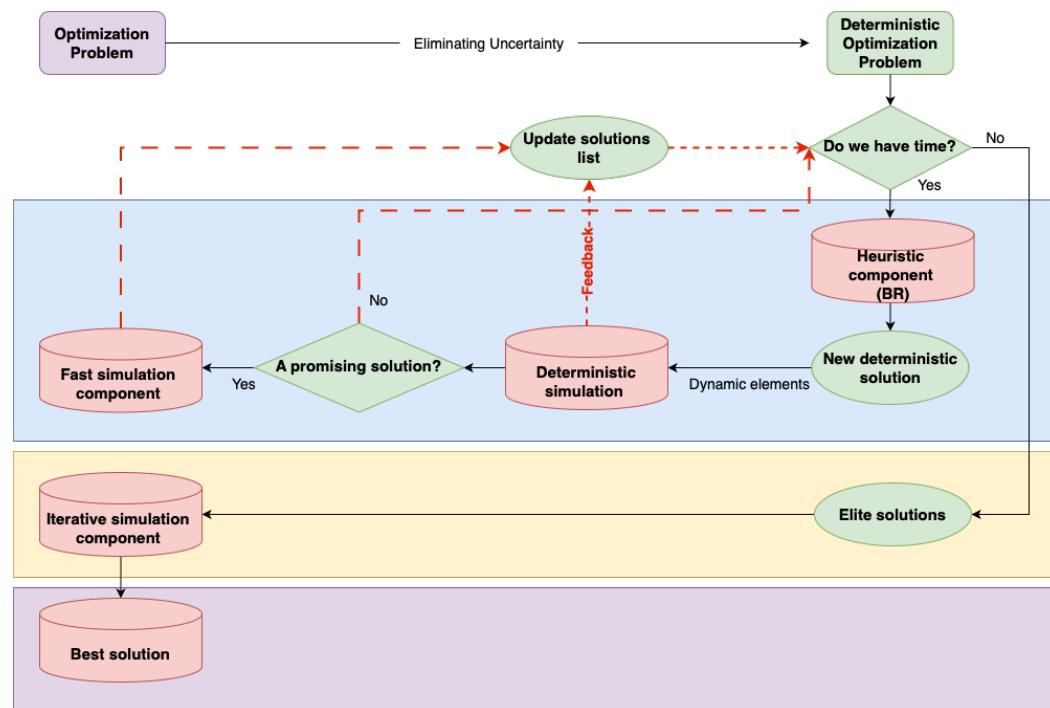


Figure 2. Schema of the heuristic utilized to solve the TOP.

Since the heuristic is randomized, a maximum time limit is set for solving each instance. The routes are merged based on their savings; instead of always choosing the best merge, it is performed in a randomized way. The merges are stored in a list and sorted based on their savings, and a probability is assigned to each one of them following a geometric distribution

with parameter β . This parameter varies between 0.1 and 0.3, as we found it provides the best solutions. Then, the routes are chosen based on that probability and merged. This allows the algorithm to find better solutions as it tries to escape local optima. With sufficient remaining time, the heuristic addresses the deterministic problem in a biased-randomized manner. If the resulting solution is promising, it is deterministically simulated several times across various dynamic environments. This component simulates dynamic elements such as traffic and weather conditions, which impact travel times, to test the robustness of the solution given by the heuristic. If the solution becomes infeasible under the dynamic conditions, its reward is set to zero. Otherwise, the solution and its associated mean reward are stored in a solution list. If the result is not promising or the short deterministic simulation is finished, the process continues iteratively until the maximum allowed time is reached. Once it is reached, the top 10 solutions from the solution list are tested in dynamic environments with a higher number of deterministic simulations. The final solution is selected from this elite group based on which performs best during these extensive tests.

The RL method, however, takes a different approach. It is built on the transformer architecture introduced by Vaswani [43], which has recently gained significant attention for solving NP-hard problems. The method follows an encoder-decoder structure, where the encoder processes variable-length input data and the decoder generates the solution. Regarding the choice of a transformer-based encoder-decoder architecture, this architecture was chosen for its ability to process input data with variable sequence lengths, making it particularly well-suited to problems like the TOP, where each problem instance can contain a different number of nodes. For a more detailed explanation, refer to Berto et al. [44], which presents state-of-the-art benchmarks, ideas, and techniques for modeling NP-hard problems such as the TOP using RL. In this case, the RL algorithm follows a constructive approach, and the flowchart of its process is shown in Figure 3.

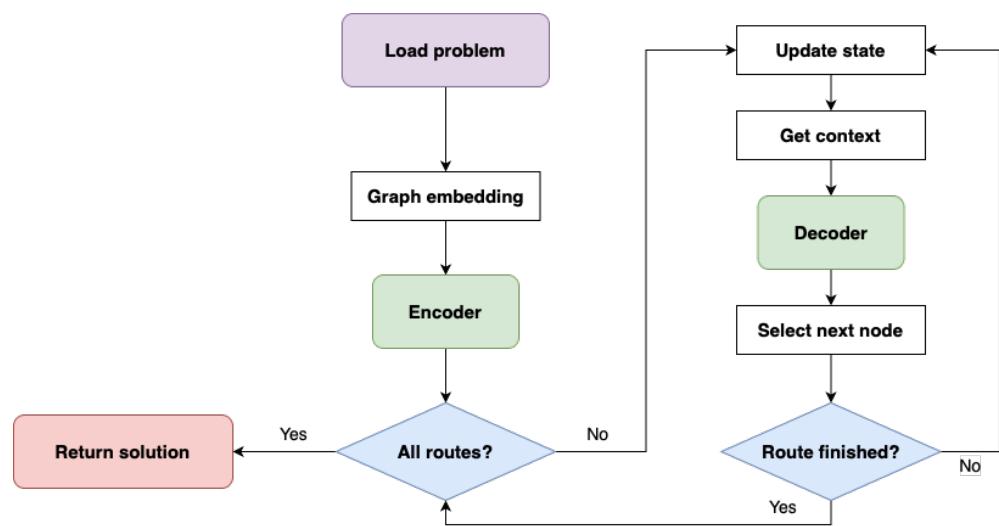


Figure 3. Schema of the RL algorithm.

First, the problem data (the position and reward of each node) are loaded into the algorithm. These data are then embedded using a linear network, with the resulting outputs passed to the encoder, which follows the architecture of the transformer encoder. The solution is constructed sequentially, route by route, and node by node. To determine the next node to visit, the algorithm updates the current state of the problem—such as the vehicle's location, the nodes available for visitation, and the current battery level. This updated state, which includes information about the graph, the vehicle's current position, the nodes already visited, and the remaining battery capacity, serves as the context provided to the model. Based on this context, the model decides which node to visit next. This process continues until the vehicle reaches the destination depot, at which point the algorithm moves on to the next route, repeating the process until all routes are completed.

For training this algorithm, the reward maximization serves as the objective function, guiding the agent's learning process to prioritize routes that provide the highest possible reward while respecting the maximum travel time. Notice that the solution generated by the RL algorithm is represented as a permutation $\pi = (\pi_1, \pi_2, \dots)$ of a subset of nodes, where the final depot can be visited more than once, but not all nodes are required to be included in the route. Policy-gradient methods learn the policy through gradient-based optimization techniques. In this framework, a stochastic policy $p(\pi|s)$ is defined to select a solution π based on a given problem s . This policy is factorized and parameterized by θ as follows:

$$p_\theta(\pi|s) = \prod_{t=1}^N p_\theta(\pi_t|s, \pi_{1:t-1}) \quad (13)$$

For training the model's policy, Williams [45] introduced a policy gradient estimator using Monte Carlo sampling, under the assumption that rewards are independent of θ . This approach is known as the REINFORCE algorithm:

$$\nabla_\theta \mathcal{L}(\theta|s) = \mathbb{E}_{p_\theta(\pi|s)}[R(\pi) \nabla_\theta \log p_\theta(\pi|s)] \quad (14)$$

However, a significant drawback of this method is its high variance, which can make model training less efficient and more unstable. To mitigate this issue, a baseline value $b(s)$ is incorporated into Equation (14), reducing variance and thereby improving training stability and overall performance:

$$\nabla_\theta \mathcal{L}(\theta|s) = \mathbb{E}_{p_\theta(\pi|s)}[(R(\pi) - b(s)) \nabla_\theta \log p_\theta(\pi|s)] \quad (15)$$

The baseline used in our case is the one proposed by Lee and Ahn [38], where problem instances and their equivalent variations, generated through augmentations, are utilized to improve the model's ability to generalize more effectively. The training phase is divided into epochs, with each epoch consisting of 2000 training steps and a batch size of 256. At the end of each epoch, the model is evaluated on 100,000 randomly generated instances to assess its performance, and if the mean reward from the validation set exceeds that of the current best model, a t -test with $\alpha = 0.05$ is performed to confirm if the new model is statistically superior. If the new model is found to be better, it is saved and used as the new baseline. The goal of the training process is for the model to continually compete against its best previous version. For the training hyperparameters, a learning rate of 5×10^{-5} was used, along with the Adam optimizer [46]. Although we experimented with higher learning rate values, the model either failed to converge or did not behave as expected during training. The training was executed on a workstation equipped with 32 GB of RAM and an NVIDIA 4060 GPU.

7. Computational Experiments

This section illustrates the experiment results after testing the algorithm described in Section 6. According to Section 5, 50 problem instances are solved, and 200 different dynamic conditions form problems for each instance, resulting in total 10,000 problems solved. The heuristic used to solve the TOP was allocated 2 s and 20 s for each instance to achieve the best result. This means that the heuristic solves each instance during 2 and 20 s, respectively. Since the heuristic is randomized, this ensures that a broader set of solutions is obtained, of which the best is returned. Those running times were chosen to show how the solution improves when letting the heuristic run for more time; moreover, the RL method takes about 0.5 s to obtain the solution, meaning that it is possible to compare computational efficiency between different solving methods. Allowing more than 20 s for the heuristic was also tested, but it did not improve performance. Additionally, the deterministic version was also solved to evaluate the impact of incorporating dynamic conditions. This version of the heuristic also runs for 2 s before returning the solution. A reward of 0 is given when the solution under dynamic conditions becomes infeasible.

Four different experiments have been run: ‘deterministic’ refers to the heuristic under deterministic conditions; ‘dynamic heu. (20 s)’ indicates the heuristic solving the problems under dynamic conditions with a time limit of 20 s for each instance; ‘dynamic heu. (2 s)’ refers to the heuristic under dynamic conditions and solving the problems within 2 s; and ‘dynamic RL’ refers to using RL to solve the problems under dynamic conditions. Notice that the deterministic heuristic approach is applied to a static version of the problem, whereas both the dynamic heuristic and the RL model solve the problem under dynamic conditions. Therefore, it is expected that the deterministic heuristic provides higher rewards, as it does not have to adjust to the uncertainties present in the dynamic scenario. We included the deterministic results intentionally to serve as a benchmark, helping to illustrate the impact that dynamic elements have on overall performance. Both the dynamic heuristic and the RL approach attempt to solve the problem without full knowledge of the dynamic conditions, which can vary significantly in real-world scenarios, especially when travel times are high enough. While better results could be achieved, this would require an estimation of the dynamic conditions beforehand.

The 50 problem instances are generated randomly: both the position of the nodes and their reward, as well as the maximum travel time allowed for each vehicle. Figure 4 shows the results obtained after running the four experiments. It is observed that the highest reward was collected by vehicles in the deterministic experiment. Both the mean and the median are the highest in this experiment compared to the other three experiments. By taking a closer look at the experiments under dynamic conditions, it is clear that the mean of the collected rewards in the dynamic RL and dynamic heu. (20 s) experiments are close and greater than those collected in dynamic heu. (2 s). The median of rewards collected in the dynamic RL experiment is the greatest. Additionally, lower variability of collected reward values is noticed in the dynamic RL experiment. This variability can be presented by the inter-quartile distance in Figure 4. The lower variability indicates greater solution reliability and ability to find consistent solutions. The greater variability in solutions found by the heuristic indicates the difficulty in finding reliable solutions.

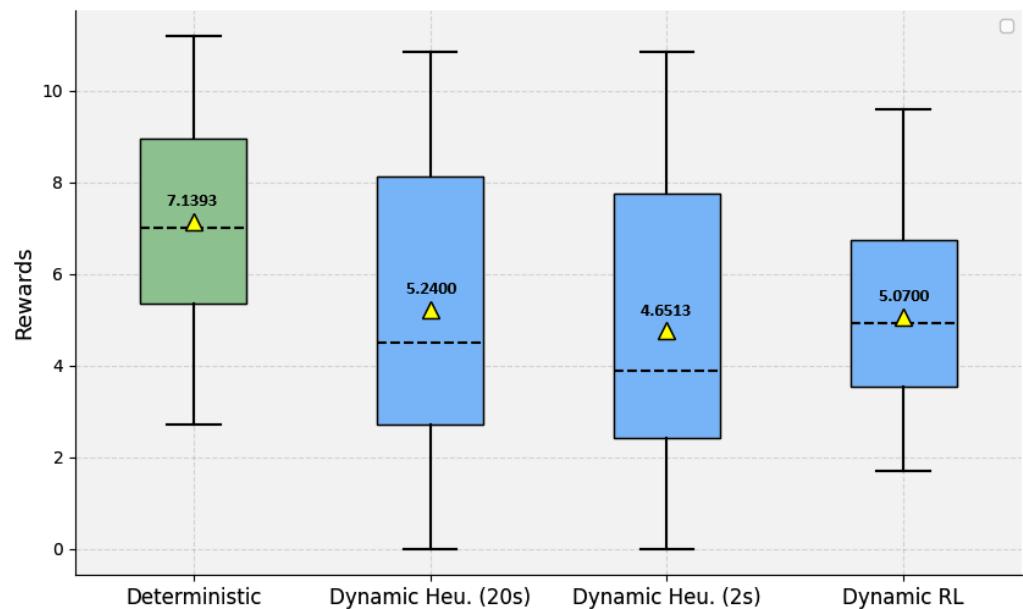


Figure 4. Distribution of results obtained in the defined four experiments.

A deeper investigation of the result is presented in Table 1. Table 1 compare between the four experiments by recording the mean of found solutions per a problem instance. The average collected reward for 21 problem instances out of the 50 instances are tabulated. These means are related to the 200 different defined dynamic condition combinations in a problem instance. Similar to the observation in Figure 4, the dynamic heu. (20 s)

experiments showed clearly better performance due to the greater exploration time compared to dynamic heu. (2 s). On average, it outperformed the dynamic RL experiment by 3.35%. Those 21 problem instances presented in Table 1 were chosen randomly from the 50 instances solved. The main goal is to show how the heuristic and RL model behave in different scenarios.

Table 1. Comparison of mean results for each problem across four solution approaches: the deterministic heuristic, the dynamic heuristic using 20 s, the dynamic heuristic using 2 s, and the RL model.

Problem	Deterministic	Dynamic Heu. (20 s)	Dynamic Heu. (2 s)	Dynamic RL
0	4.2553	3.1504	2.1379	3.5322
1	8.7905	8.1820	4.6382	5.4216
2	4.0850	2.6747	1.8263	2.8274
3	6.4999	6.3844	3.2809	4.8508
4	8.3060	6.5841	6.4133	6.4047
5	5.8537	3.8344	3.7612	3.1818
6	9.0080	9.0080	9.0080	7.7950
7	11.1873	10.7949	10.7949	8.4732
8	6.6231	5.1850	5.1850	4.9351
9	4.7175	2.0463	2.0240	3.4620
10	9.2883	5.3923	0.0000	6.1189
11	6.6812	0.0694	0.1005	3.8277
12	5.3447	3.1942	3.1942	3.3079
13	8.8668	8.8668	8.8668	7.1570
14	2.7068	1.5510	1.5510	2.0699
15	5.3801	5.0151	5.0151	4.3966
16	7.0243	3.3182	3.2398	4.1747
17	4.0835	2.4231	2.4231	2.9224
18	5.8006	2.0723	2.0723	3.8718
19	7.7680	4.4660	3.0630	5.0875
20	4.7668	4.4149	3.5105	3.9444
:	:	:	:	:
Avg	7.1393	5.2400	4.6513	5.0700

The solution found for the deterministic version of the problem instances is considered as the upper bound. With introduced dynamic conditions, the vehicle travel time increases and reduces the possibility to collect all rewards as in the deterministic version of the problem. Depending on the severity of the weather and traffic conditions, the collected rewards are affected and reduced accordingly. In some problem instances, such as instance 6 and 13 in Table 1, the heuristic was successful in finding the greatest reward under the dynamic conditions, while the RL model was not able to find. In other instances, such as instances 1 and 9, the RL model was successful in solutions with rewards greater than those found by the heuristic under the dynamic conditions. In Figure 5, the training validation reward of the RL model is illustrated. As expected, the model shows a rapid improvement during the initial epochs. However, the rate of improvement significantly decreases in later epochs. The red dots show the epochs when the model improves, at which point the new baseline is saved.

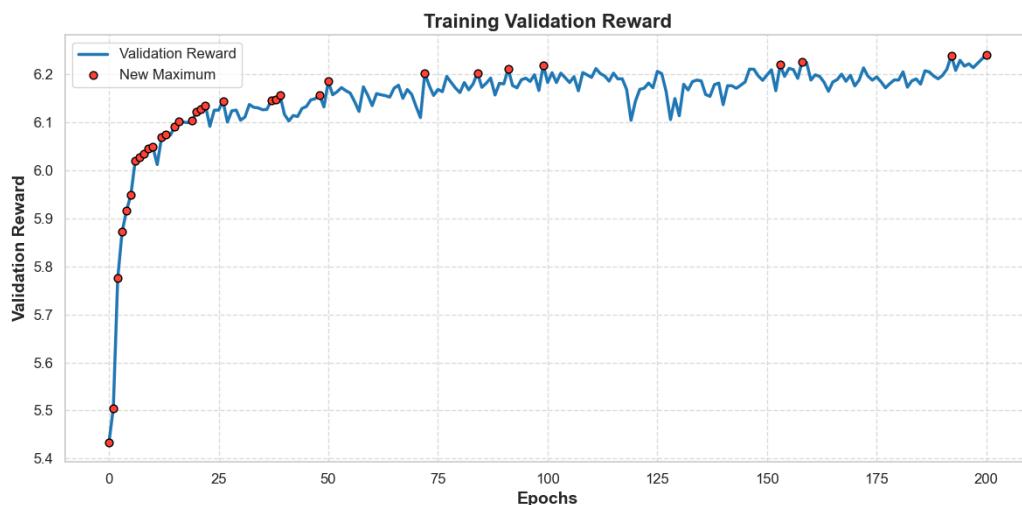


Figure 5. Validation reward at each epoch in the training phase.

It is also important to highlight the time required to train the model. In this case, training for 200 epochs took approximately 16 h. Additionally, a significant amount of training data is needed, which can be challenging to generate, particularly for more complex problems that simulate intricate real-world conditions and are hard to validate. The scalability of the RL model also requires attention, as varying the number of nodes from those used during training may result in sub-optimal performance. According to the results, the training of the RL model is challenging and demands resources. Once the RL model is trained, it provides promising and reliable solutions. The solution reliability is presented by the lower variability compared to the other approaches (Figure 4).

8. Conclusions

This paper presents a dual approach to solve the dynamic TOP with EVs, constrained by battery driving range and impacted by real-time dynamic conditions. To address this challenge, we propose two independent methodologies: a heuristic-based approach for rapid solution generation and a reinforcement learning approach for adaptive decision-making. By applying these distinct methods, we are able to explore the strengths and weaknesses of each in scenarios where dynamic factors like road congestion, battery status, and travel times fluctuate continuously.

The heuristic approach efficiently produces initial solutions, particularly in relatively stable environments, providing a fast and practical option for route planning. In contrast, the reinforcement learning approach excels in more dynamic environments, learning to adapt decisions as conditions evolve, and consistently achieving higher-quality solutions by optimizing both the number of nodes visited and energy efficiency. The computational experiments reveal that while the heuristic approach performs well under deterministic conditions, it might lack the flexibility needed to adapt in highly variable scenarios. The reinforcement learning approach, however, demonstrates its capacity to incorporate dynamic elements, offering robust and reliable solutions even under significant uncertainty. Although RL experiment showed a great performance in this study in solving the TOP, the training of the RL model is computationally demanding.

In this study, the RL approach is used to solve TOP as an example of a last-mile delivery problem, involving dynamic components. The dynamic travel time represents one of real-time aspects encountered in optimization problems in addition to stochastic uncertainty. The RL approach showed the ability to handle the dynamic conditions and recommended solutions after training the RL model to handle such problems. Similar applications could be found in other last-mile delivery problems, especially those reflecting real-time and real-world problems.

Our future work will focus on introducing uncertainty along with the dynamic conditions to the TOP problem, especially in regard to battery management and duration. Additionally, investigating approaches for speeding up training of the reinforcement learning model is also a line to consider. The scalability of the solving approach is an issue to further investigate and evaluate in future works. In real-life problems, the batteries life variability is one of aspects to be considered. Accordingly, the future work can expand the problem definition and consider batteries life as well.

Author Contributions: Conceptualization, A.A.J.; methodology, A.G., V.T., C.S. and M.A.; software, A.G.; validation, M.A. and C.S.; writing—original draft preparation, V.T., A.G. and C.S.; writing—review and editing, V.T., M.A. and A.A.J.; supervision, V.T. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding: The present work has been partially funded by the Spanish Government (IA4TES project ‘Artificial Intelligence for Sustainable Energy Transition’), the Spanish Ministry of Science-AEI (PID2022-138860NB-I00 and RED2022-134703-T), as well as the European Commission (SUN HORIZON-CL4-2022-HUMAN-01-14-101092612, AIDEAS HORIZON-CL4-2021-TWIN-TRANSITION-01-07-101057294).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BEV	battery electric vehicle
DRL	deep reinforcement learning
EV	electric vehicle
FCEV	fuel cell electric vehicles
HEV	hybrid electric vehicle
PHEV	plug-in hybrid electric vehicle
RL	reinforcement learning
UAV	unmanned aerial vehicle

References

1. Puzicha, A.; Buchholz, P. Dynamic Mission Control for Decentralized Mobile Robot Swarms. In Proceedings of the 2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), Sevilla, Spain, 8–10 November 2022; pp. 257–263.
2. Rabe, M.; Poeting, M.; Klueter, A. Evaluating the Benefits of Collaborative Distribution with Supply Chain Simulation. In *Food Supply Chains in Cities: Modern Tools for Circularity and Sustainability*; Palgrave Macmillan: Cham, Switzerland, 2020; pp. 69–100.
3. Poeting, M.; Prell, B.; Rabe, M.; Uhlig, T.; Wenzel, S. Considering energy-related factors in the simulation of logistics systems. In Proceedings of the 2019 Winter Simulation Conference (WSC), National Harbor, MD, USA, 8–11 December 2019; pp. 1849–1858.
4. Poeting, M.; Schaudt, S.; Clausen, U. A comprehensive case study in last-mile delivery concepts for parcel robots. In Proceedings of the 2019 Winter Simulation Conference (WSC), National Harbor, MD, USA, 8–11 December 2019; pp. 1779–1788.
5. Khan, A.; Zhang, J.; Ahmad, S.; Memon, S.; Qureshi, H.A.; Ishfaq, M. Dynamic positioning and energy-efficient path planning for disaster scenarios in 5G-assisted multi-UAV environments. *Electronics* **2022**, *11*, 2197. [[CrossRef](#)]
6. Khan, S.I.; Qadir, Z.; Munawar, H.S.; Nayak, S.R.; Budati, A.K.; Verma, K.D.; Prakash, D. UAVs path planning architecture for effective medical emergency response in future networks. *Phys. Commun.* **2021**, *47*, 101337. [[CrossRef](#)]
7. Golden, B.; Levy, L.; Vohra, R. The Orienteering Problem. *Nav. Res. Logist.* **1987**, *34*, 307–318. [[CrossRef](#)]
8. Panadero, J.; Currie, C.; Juan, A.A.; Bayliss, C. Maximizing Reward from a Team of Surveillance Drones under Uncertainty Conditions: A Simheuristic Approach. *Eur. J. Ind. Eng.* **2020**, *14*, 1–23. [[CrossRef](#)]
9. Sebai, M.; Rejeb, L.; Denden, M.A.; Amor, Y.; Baati, L.; Said, L.B. Optimal electric vehicles route planning with traffic flow prediction and real-time traffic incidents. *Int. J. Electr. Comput. Eng. Res.* **2022**, *2*, 1–12. [[CrossRef](#)]
10. Peng, Z.; Li, B.; Chen, X.; Wu, J. Online route planning for UAV based on model predictive control and particle swarm optimization algorithm. In Proceedings of the 10th World Congress on Intelligent Control and Automation, Beijing, China, 6–8 July 2012; pp. 397–401.
11. Juan, A.A.; Keenan, P.; Martí, R.; McGarragh, S.; Panadero, J.; Carroll, P.; Oliva, D. A review of the role of heuristics in stochastic optimisation: From metaheuristics to learnheuristics. *Ann. Oper. Res.* **2023**, *320*, 831–861. [[CrossRef](#)]
12. Szepesvári, C. *Algorithms for Reinforcement Learning*; Springer Nature: Cham, Switzerland, 2022.

13. Zhao, X.; Ke, Y.; Zuo, J.; Xiong, W.; Wu, P. Evaluation of sustainable transport research in 2000–2019. *J. Clean. Prod.* **2020**, *256*, 120404. [[CrossRef](#)]
14. IPCC. *Climate Change 2022: Mitigation of Climate Change*; Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; Cambridge University Press, Cambridge, UK; New York, NY, USA, 2022.
15. IEA. *Global EV Outlook 2023*; International Energy Agency: Paris, France, 2023.
16. Alanazi, F. Electric vehicles: Benefits, challenges, and potential solutions for widespread adaptation. *Appl. Sci.* **2023**, *13*, 6016. [[CrossRef](#)]
17. Corradi, C.; Sica, E.; Morone, P. What drives electric vehicle adoption? Insights from a systematic review on European transport actors and behaviours. *Energy Res. Soc. Sci.* **2023**, *95*, 102908. [[CrossRef](#)]
18. Ntombela, M.; Musasa, K.; Moloi, K. A comprehensive review for battery electric vehicles (BEV) drive circuits technology, operations, and challenges. *World Electr. Veh. J.* **2023**, *14*, 195. [[CrossRef](#)]
19. Cao, Y.; Yao, M.; Sun, X. An overview of modelling and energy management strategies for hybrid electric vehicles. *Appl. Sci.* **2023**, *13*, 5947. [[CrossRef](#)]
20. Martinez, C.M.; Hu, X.; Cao, D.; Velenis, E.; Gao, B.; Wellers, M. Energy management in plug-in hybrid electric vehicles: Recent progress and a connected vehicles perspective. *IEEE Trans. Veh. Technol.* **2016**, *66*, 4534–4549. [[CrossRef](#)]
21. König, A.; Nicoletti, L.; Schröder, D.; Wolff, S.; Waclaw, A.; Lienkamp, M. An overview of parameter and cost for battery electric vehicles. *World Electr. Veh. J.* **2021**, *12*, 21. [[CrossRef](#)]
22. Waseem, M.; Amir, M.; Lakshmi, G.S.; Harivardhagini, S.; Ahmad, M. Fuel cell-based hybrid electric vehicles: An integrated review of current status, key challenges, recommended policies, and future prospects. *Green Energy Intell. Transp.* **2023**, *2*, 100121. [[CrossRef](#)]
23. Pramuanjaroenkij, A.; Kakaç, S. The fuel cell electric vehicles: The highlight review. *Int. J. Hydrogen Energy* **2023**, *48*, 9401–9425. [[CrossRef](#)]
24. Khan, A.; Yaqub, S.; Ali, M.; Ahmad, A.W.; Nazir, H.; Khalid, H.A.; Iqbal, N.; Said, Z.; Sopian, K. A state-of-the-art review on heating and cooling of lithium-ion batteries for electric vehicles. *J. Energy Storage* **2024**, *76*, 109852. [[CrossRef](#)]
25. Duan, J.; Tang, X.; Dai, H.; Yang, Y.; Wu, W.; Wei, X.; Huang, Y. Building safe lithium-ion batteries for electric vehicles: A review. *Electrochem. Energy Rev.* **2020**, *3*, 1–42. [[CrossRef](#)]
26. Ahmad, F.; Saad Alam, M.; Saad Alsaidan, I.; Shariff, S.M. Battery swapping station for electric vehicles: Opportunities and challenges. *IET Smart Grid* **2020**, *3*, 280–286. [[CrossRef](#)]
27. Li, W.; Stanula, P.; Egede, P.; Kara, S.; Herrmann, C. Determining the main factors influencing the energy consumption of electric vehicles in the usage phase. *Procedia CIRP* **2016**, *48*, 352–357. [[CrossRef](#)]
28. Bi, J.; Wang, Y.; Zhang, J. A data-based model for driving distance estimation of battery electric logistics vehicles. *EURASIP J. Wirel. Commun. Netw.* **2018**, *2018*, 251. [[CrossRef](#)]
29. Togun, H.; Aljibori, H.S.S.; Biswas, N.; Mohammed, H.I.; Sadeq, A.M.; Rashid, F.L.; Abdulrazzaq, T.; Zearah, S.A. A critical review on the efficient cooling strategy of batteries of electric vehicles: Advances, challenges, future perspectives. *Renew. Sustain. Energy Rev.* **2024**, *203*, 114732. [[CrossRef](#)]
30. Brenna, M.; Foiadelli, F.; Leone, C.; Longo, M. Electric Vehicles Charging Technology Review and Optimal Size Estimation. *J. Electr. Eng. Technol.* **2020**, *15*, 2539–2552. [[CrossRef](#)]
31. Yong, J.Y.; Ramachandaramurthy, V.K.; Tan, K.M.; Mithulananthan, N. A Review on the State-of-the-Art Technologies of Electric Vehicle, Its Impacts and Prospects. *Renew. Sustain. Energy Rev.* **2015**, *49*, 365–385. [[CrossRef](#)]
32. Montoya, A.; Guéret, C.; Mendoza, J.E.; Villegas, J.G. The electric vehicle routing problem with nonlinear charging function. *Transp. Res. Part Methodol.* **2017**, *103*, 87–110. [[CrossRef](#)]
33. Zografos, K.G.; Androulatsopoulos, K.N.; Vasilakis, G.M. A real-time decision support system for roadway network incident response logistics. *Transp. Res. Part C Emerg. Technol.* **2002**, *10*, 1–18. [[CrossRef](#)]
34. Xu, Y.; Fang, M.; Chen, L.; Xu, G.; Du, Y.; Zhang, C. Reinforcement learning with multiple relational attention for solving vehicle routing problems. *IEEE Trans. Cybern.* **2021**, *52*, 11107–11120. [[CrossRef](#)] [[PubMed](#)]
35. Vincent, F.Y.; Salsabila, N.Y.; Lin, S.W.; Gunawan, A. Simulated annealing with reinforcement learning for the set team orienteering problem with time windows. *Expert Syst. Appl.* **2024**, *238*, 121996.
36. Panadero, J.; Juan, A.A.; Ghorbani, E.; Faulin, J.; Pagès-Bernaus, A. Solving the stochastic team orienteering problem: Comparing simheuristics with the sample average approximation method. *Int. Trans. Oper. Res.* **2024**, *31*, 3036–3060. [[CrossRef](#)]
37. Sankaran, P. *Deep Reinforcement Learning and Hybrid Approaches to Solve Multi-Vehicle Combinatorial Optimization Problems*; Rochester Institute of Technology: Rochester, NY, USA, 2023.
38. Lee, D.H.; Ahn, J. Multi-start team orienteering problem for UAS mission re-planning with data-efficient deep reinforcement learning. *Appl. Intell.* **2024**, *54*, 4467–4489. [[CrossRef](#)]
39. Wang, Y.; Zhou, J.; Sun, Y.; Fan, J.; Wang, Z.; Wang, H. Collaborative multidepot electric vehicle routing problem with time windows and shared charging stations. *Expert Syst. Appl.* **2023**, *219*, 119654. [[CrossRef](#)]
40. Sánchez, D.G.; Tabares, A.; Faria, L.T.; Rivera, J.C.; Franco, J.F. A clustering approach for the optimal siting of recharging stations in the electric vehicle routing problem with time windows. *Energies* **2022**, *15*, 2372. [[CrossRef](#)]
41. Juan, A.A.; Marugan, C.A.; Ahsini, Y.; Fornes, R.; Panadero, J.; Martin, X.A. Using Reinforcement Learning to Solve a Dynamic Orienteering Problem with Random Rewards Affected by the Battery Status. *Batteries* **2023**, *9*, 416. [[CrossRef](#)]

42. Wang, R.; Liu, W.; Li, K.; Zhang, T.; Wang, L.; Xu, X. Solving Orienteering Problems by Hybridizing Evolutionary Algorithm and Deep Reinforcement Learning. *IEEE Trans. Artif. Intell.* **2024**, *5*, 5493–5508. [[CrossRef](#)]
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
44. Berto, F.; Hua, C.; Park, J.; Luttmann, L.; Ma, Y.; Bu, F.; Wang, J.; Ye, H.; Kim, M.; Choi, S.; et al. Rl4co: An extensive reinforcement learning for combinatorial optimization benchmark. *arXiv* **2023**, arXiv:2306.17100.
45. Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256. [[CrossRef](#)]
46. Ruthotto, L.; Haber, E. An introduction to deep generative modeling. *GAMM-Mitteilungen* **2021**, *44*, e202100008. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.