# Optimizing EV Charging Routes with Reinforcement Learning (Proximal Policy Optimization)

*Integrating tariff heterogeneity, nonlinear charging curves, and traffic simulation for 9 Boroughs in Inner London*

Hatim Shaherawala[1[21054059]], Vishesh Jakhar[1[24058353]], and Divykumar Patel[1[24058610]]

[1] University of The West of England, Bristol BS16 1QY, United Kingdom

**Abstract:** The growing adoption of electric vehicles (EVs) poses new challenges for urban mobility, particularly in cities with dense infrastructure and diverse charging tariffs. This thesis develops a reinforcement learning framework for optimising EV charging routes in Inner London, integrating tariff heterogeneity, nonlinear charging curves, and traffic congestion. A custom Gym–SUMO environment was implemented, and Proximal Policy Optimisation (PPO) was applied to minimise cost, time, and hybrid objectives. Evaluation shows that cost-based agents achieved near-perfect success with stable state-of-charge outcomes, while time-based agents systematically failed due to infeasibility. The hybrid formulation collapsed into cost minimisation, highlighting the importance of reward design and parameter calibration. The results demonstrate that reinforcement learning can inform driver-centric EV charging strategies and identify future directions in dynamic pricing, multi-objective optimisation, and environment realism.

**Keywords:** Electric vehicles · Charging optimisation · Reinforcement learning · Proximal Policy Optimisation (PPO) · SUMO · Multi-objective optimisation · Urban mobility

# 1 Introduction

The rapid adoption of electric vehicles (EVs) is reshaping transport and energy systems, driven by ambitious decarbonisation targets and urban air quality policies [8]. In London, this transition has been supported by an expansion of charging infrastructure [10], yet challenges persist. The charging landscape remains fragmented across multiple operators, with uneven station distribution [9] and tariffs that combine per-kWh energy rates, session fees, idle penalties, and membership discounts [13, 32]. For drivers, key concerns include minimising costs, reducing delays, and ensuring trip feasibility, while operators and regulators focus on infrastructure adequacy and grid stability. Existing optimisation efforts have often prioritised the grid perspective [1–3] but struggle to capture the complex, user-centred realities of urban charging.

Several gaps in the literature limit the applicability of current approaches to dense metropolitan contexts like Inner London. Many studies assume uniform or simplified tariffs, overlooking operator heterogeneity, idle fees, and subscription models [13–15, 32]. Charging dynamics are frequently simplified as well: constant or generic rates are used in place of vehicle-specific nonlinear charging curves [17, 33, 34]. Hybrid objectives balancing cost and time remain underexplored; weighted-sum and Pareto methods are sometimes applied, but their calibration is often ad hoc [35–37]. Reinforcement learning (RL) has shown promise in this domain [5–7, 21, 22], yet many implementations rely on oversimplified environments that neglect congestion, unrealistic station cycling, or other real-world constraints [16, 25, 41]. Reward design also poses difficulties: prior work often fixes objective priorities or under-utilises shaping and penalties [38–40]. Collectively, these gaps indicate that a new, driver-centric approach is needed. One that captures real-world variability in tariffs, charging behaviour, and urban constraints.

This thesis addresses these challenges by applying Proximal Policy Optimisation (PPO) [19] to model EV charging decisions in Inner London. Using structured datasets, the study integrates operator-specific tariffs, detailed station metadata, and vehicle charging curves to simulate realistic charging costs and times. A custom RL environment was built using SUMO for traffic modelling and OpenAI Gym for training, with rules to prevent implausible behaviours such as repeated station cycling. Reward functions were designed for cost, time, and hybrid objectives, incorporating potential-based shaping and feasibility penalties. The methodology emphasises reproducibility through structured data inputs, fixed random seeds, and consistent evaluation protocols.

The research is guided by four questions:

Q1: How can reinforcement learning minimise total trip cost while ensuring state of charge (SoC) remains above a safe reserve threshold?

Q2: How can reinforcement learning minimise total trip duration while maintaining SoC for journey completion?

Q3: How can a hybrid reinforcement learning framework balance cost and time objectives under varying operational constraints?

Q4: How do cost-based, time-based, and hybrid optimisation objectives differ in shaping charging behaviour and route selection, and what trade-offs emerge between them?

By addressing these questions, the study contributes a reinforcement learning framework that captures the technical and behavioural complexities of EV charging in Inner London.

## 2 Literature Review

Before reinforcement learning gained prominence, EV charging optimisation was approached with deterministic and heuristic methods. Mixed-integer linear programming (MILP) can yield optimal solutions under fixed assumptions [16, 17] but scales poorly for real-time, city-wide applications as the number of vehicles and stations grows. Metaheuristic techniques (e.g. genetic algorithms, particle swarm optimisation) improve scalability [20], yet early electric vehicle routing formulations still overlooked critical dynamics like traffic congestion, waiting queues, and nonlinear charging rates [16]. These limitations prompted a shift towards adaptive, data-driven methods, with RL offering a compelling alternative.

Many prior studies oversimplified key technical factors. Battery charging is nonlinear, power input tapers as the battery fills. Yet models often assume a constant rate, underestimating charging duration [16, 17]. Vehicle-specific differences are usually ignored due to limited data availability [33]; only recently have large datasets revealed broad heterogeneity in EV charging behaviour across models [34]. Real-world pricing is likewise complex, operators impose varying energy rates, session fees, idle penalties, and membership discounts [13, 32]. However, much of the earlier optimisation literature assumed uniform or flat pricing schemes [13–15, 32], which can yield unrealistic recommendations. Incorporating these tariff nuances (for example, discounted member-only rates vs. higher guest prices, or per-minute idle fees) can significantly change optimal charging decisions [15, 37]. Any practical framework must therefore reflect such tariff heterogeneity and realistic charging dynamics, especially in dense urban settings.

RL offers the ability to learn adaptive policies through interaction with the environment, something static optimisations cannot achieve. Applications of RL in energy and transportation show it can reduce costs and improve operational efficiency [12, 21, 22]. Early work on EV charging used value-based RL (e.g. Deep Q-Networks) to coordinate charging and reduce user cost peaks [5, 6], but the discrete action space in those approaches was not suitable for continuous routing decisions. Policy-gradient methods like Proximal Policy Optimization (PPO) can handle continuous actions [7] and are known for stable training; indeed, PPO has demonstrated strong performance on complex or multi-objective tasks in simulations [23]. Despite this promise, few studies to date have applied PPO for individual driver routing in dense cities, where variable tariffs, nonlinear charging curves, and traffic congestion all intersect. Other advanced RL algorithms (e.g., DDPG and SAC) have also been tested for EV charging control [31], but they similarly have not been deployed at the driver-route level in a congested urban context.

A realistic simulation environment is essential for training RL agents effectively. Many earlier studies made simplifying assumptions such as deterministic travel times or unconstrained charger availability, ignoring urban traffic and station queues [16]. In contrast, recent works stress simulation realism by integrating traffic models and operational constraints. Using tools like OpenAI Gym with the SUMO traffic simulator allows an agent to experience time-varying congestion and competition for charging stalls during training [41]. This added fidelity prevents overly optimistic strategies and yields policies that, for example, learn to avoid heavily congested routes or anticipate waiting times at busy stations, behaviours crucial for real-world feasibility.

Designing an effective reward function is another challenge when multiple objectives must be balanced. Past works often focused on a single goal (either cost or time) or used a fixed weighted sum of both [21, 22]. Some researchers even convert time into an equivalent monetary cost via a value-of-time

factor to merge objectives into one metric [35, 37]. Alternatively, evolutionary and Pareto-based methods have been explored to avoid preset weights and generate a spectrum of cost–time trade-off solutions [36]. In practice, careful reward shaping is needed. One proven technique is potential-based shaping, which adds an extra incentive (e.g. rewarding reductions in remaining travel time or distance) without changing the optimal policy [38]. Strong penalty terms are also common: for instance, a very large negative reward if the vehicle depletes its battery en route trains the agent to avoid infeasible paths [39]. Effective use of shaping and penalties significantly improves learning efficiency and policy realism in EV charging simulations [38, 39]. Notably, a few multi-objective RL approaches for EV charging have been proposed in recent years [29, 30], but these remain limited in scope and have yet to tackle the full real-world complexity (heterogeneous prices, nonlinear charging, traffic congestion) that this study addresses.

In summary, the existing literature highlights the need for a driver-centric EV charging strategy that can adapt to complex urban conditions. No prior work fully integrates fine-grained tariff differences, realistic charging curves, and dynamic traffic into a single optimisation framework. The present study therefore applies PPO in a realistic Inner London simulation, explicitly integrating these real-world factors into a unified RL approach for EV routing and charging.

## 3 Methodology

### 3.1 Data Preparation

Four key data sources were compiled to drive the simulation:

**1. Tariffs**: Real world pricing data were collected from multiple charging network operators [44 -58] (see Fig. 1) and standardized into a unified format (converted into £/kWh units) to enable accurate cost calculations.

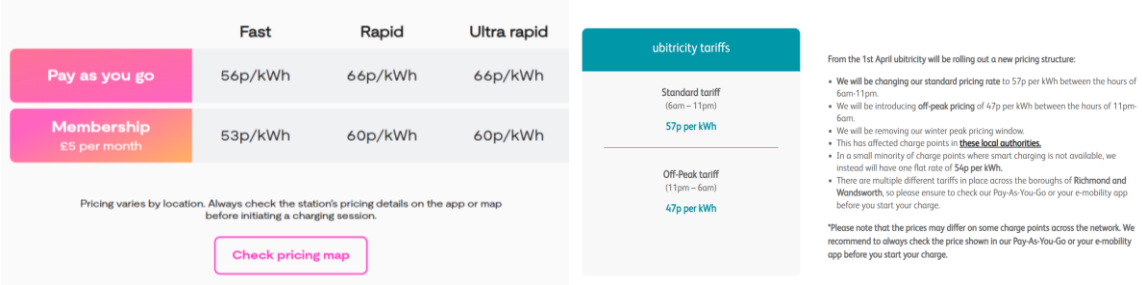


**Figure 1: Raw Tariff Information Examples**

Tariffs were split into three structured datasets. This ensured realistic and provider-specific cost estimation in the simulation environment (see Fig. 2).

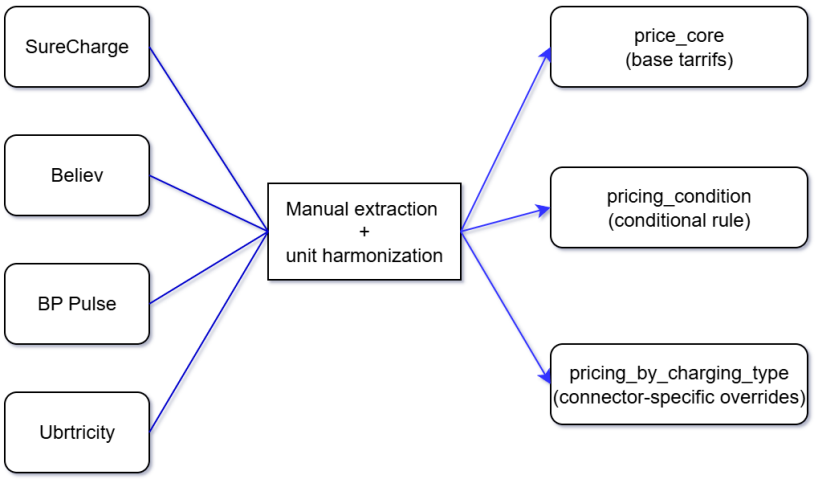


**Figure 2 Conversion Process**

**2. Stations:** The UK National Chargepoint Registry [42] provided station metadata, which were cleaned and split into station-level metadata (e.g. location, operator, number of connectors) and connector-level attributes (power output, plug type, tariff info) for use in the simulation (see Fig. 3). For geographic scope, stations were filtered to Inner London, restricted to nine boroughs (Camden, City of London, Hackney, Hammersmith and Fulham, Islington, Kensington and Chelsea, Lambeth, Tower Hamlets, Westminster).

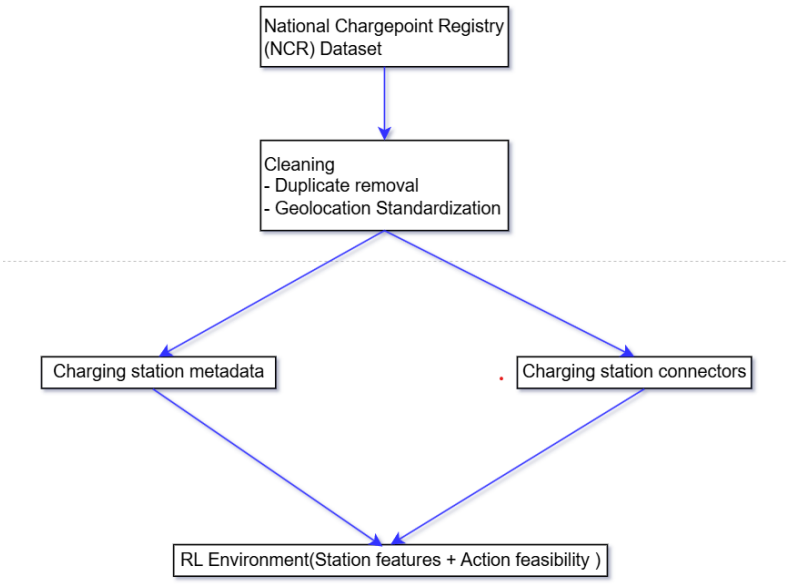


**Figure 3 Splitting the Charging Data**

**3. Vehicles:** Specifications such as battery capacity and charging curves were sourced from the Open EV Data project [43], enabling realistic modelling of charging times.

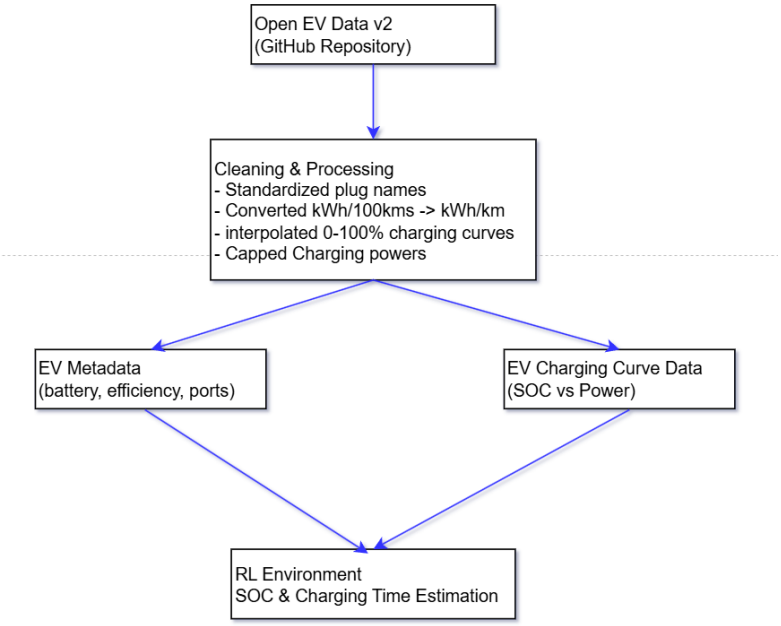


**Figure 4 Splitting the EV Data**

**4. User trips:** A set of simulated trip scenarios was generated, each specifying an origin, destination, and start time along with an initial SoC (with a set reserve margin), a particular EV model, the driver's membership status (for tariff effects), and the chosen optimisation objective (cost, time, or hybrid).

These structured datasets ensured that the RL environment had access to rich, realistic information on costs, infrastructure, vehicle behaviour and driver preferences.

### 3.2 Environment Design

A custom OpenAI Gym environment was implemented with SUMO as the traffic simulator [25]. At each decision step, the agent observes the current state and chooses either to continue driving or divert to a charging station.

The environment is a Markov Decision Process (MDP) defined in Eq. (1).

$$\boldsymbol{M} = (\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{P}, \boldsymbol{R}, \boldsymbol{\gamma}) \quad \textbf{(1)}$$

Driving energy use and state-of-charge (SoC) updates follow Eq. (2), with travel times obtained from SUMO or congestion-scaled estimates Eq. (3). Such SoC and consumption formulations are widely adopted in EV routing problems [16, 17].

$$\Delta \boldsymbol{E}_{\text{drive}} = \boldsymbol{c}_{\text{veh}} \cdot \Delta \boldsymbol{d}, \quad \text{SoC}_{t+1} = \text{SoC}_{t} - \frac{\Delta E_{\text{drive}}}{B} \quad \textbf{(2)}$$

$$\Delta \boldsymbol{t}_{\text{drive}} = \boldsymbol{t}_{\text{SUMO}}(\boldsymbol{x}_t \rightarrow \boldsymbol{x}_{t+1}) \quad \textbf{(3)}$$

Charging respects connector limits and vehicle-specific charging curves (Eq. 4), with a fixed 3-minute session overhead. This nonlinear tapering behaviour reflects empirical charging models validated in EV routing studies [16, 17].

$$P_{\text{eff}}(\text{SoC}_t) = \min\{P_{\text{conn}}, P_{\text{curve}}(\text{SoC}_t)\} \quad \textbf{(4)}$$

$$\Delta E_{\text{chg}} = \eta\, P_{\text{eff}}(\text{SoC}_t)\, \Delta t_{\text{chg}}, \quad \text{SoC}_{t+1} = \min\left\{1,\ \text{SoC}_t + \frac{\Delta E_{\text{chg}}}{B}\right\} \quad \textbf{(4)}$$

$$\Delta t_{\text{chg,tot}} = \Delta t_{\text{chg}} + \tau_{\text{overhead}}, \quad \tau_{\text{overhead}} = 180\ \text{s} \quad \textbf{(4)}$$

Episodes terminate when the destination is reached with SoC above reserve, or when depletion occurs. Invalid actions (e.g. selecting unreachable or repeated stations, charging before cooldown, exceeding trip charge limits) incur penalties and may trigger termination.

Traffic dynamics can be run in full SUMO microsimulation or in a lightweight congestion mode, applying peak-hour multipliers (1.6 for AM, 1.5 for PM) to represent travel delays.

The environment supports three optimisation objectives (time, cost, hybrid). A global preference is set by default, but if enabled, the objective can be overridden per trip to allow heterogeneous agent behaviour. When operating in pure time mode, stabilization heuristics reduce per-minute penalties, boost completion rewards, amplify shaping, and halve charging overheads to prevent value-function collapse.

Design trade-offs are summarised in **Table 1**.

**Table 1: Environment Design Trade-Offs**

| Aspect | Real-world baseline | Simulation choice | Rationale |
|---|---|---|---|
| **Initial SoC** | Often >50% | Sampled 10–30% | Ensures charging is frequently required, avoids trivial trips |
| **Trip lengths** | Many <10 km | Calibrated 12–25 km | 30–60% of trips require ≥1 charge, providing a learning signal |
| **Charging** | Overheads vary | Fixed 3-min per session | Penalises “nibbling” charges and station hopping |
| **Traffic** | Complex congestion patterns | SUMO microsim (vs. constant speed or multipliers) | Provides realistic congestion while keeping simulation deterministic |
| **Station use** | Drivers may revisit stations | No repeats, cooldowns, max charges | Prevents unrealistic cycling behaviour |
| **Variability** | High randomness in trips | Fixed seeds (environment + training) | Enables reproducibility and controlled comparisons |

### 3.3 Reward Design

Three optimisation objectives were supported: cost, time, and hybrid (via a value-of-time conversion). Per-step rewards are given in Eqs. (5–7).

**Cost objective (per step)**

$$r_t^{\text{cost}} = -C_t, \quad C_t = \underbrace{p_t \, \Delta E_{\text{chg}}}_{\text{energy cost}} + \text{session}_t + \text{idle}_t \quad \textbf{(5)}$$

**Time objective (per step)**

$$r_t^{\text{time}} = -\Delta t_t \quad \textbf{(6)}$$

**Hybrid objective with value-of-time**

$$r_t^{\text{hyb}} = -(C_t + \alpha \, \Delta t_t), \quad \alpha = 0.05 \, £/\text{min} \quad \textbf{(7)}$$

Potential-based shaping (Eq. 8) was used to provide incremental feedback based on reductions in estimated arrival time. It is a proven reinforcement learning technique to accelerate convergence without altering the optimal policy [38].

**Potential-based shaping and penalties**

$$F(s_t, s_{t+1}) = \gamma \, \Phi(s_{t+1}) - \Phi(s_t), \quad \Phi(s) = -\kappa \, \text{ETA}(s) \quad \textbf{(8)}$$

$$r_t = r_t^{(\cdot)} + F(s_t, s_{t+1}) - \epsilon \, \mathbf{1}\{\text{invalid}\} - \xi \, \mathbf{1}\{\text{micro}\} \quad \textbf{(8)}$$

$$r_T \leftarrow r_T + R_{\text{succ}} \, \mathbf{1}\{\text{success}\} - R_{\text{fail}} \, \mathbf{1}\{\text{stranded}\} \quad \textbf{(8)}$$

Charging steps were always net-negative, as both energy and time costs are incurred. Terminal rewards penalised stranding while rewarding successful completion. Constraint violations, such as repeated station visits or cooldown breaches, introduced additional penalties to guide exploration. Similar feasibility constraints and heavy stranding penalties were seen in a study [39].

The specific values for terminal rewards, penalties, and shaping parameters, along with their rationale, are summarised in Table 2.

**Table 2: Summary of Design Choices**

| Design Choice | Implementation (Value) | Rationale |
|---|---|---|
| **Success / failure** | +50 (completion), –200 (stranding) | Encourages feasibility and strongly penalises infeasible trips |
| **Infeasible action** | –2 | Discourages wasted steps while keeping exploration possible |
| **Repeat station penalty** | –5 | Prevents unrealistic cycling behaviour |

| | | |
|---|---|---|
| **Cooldown violation** | –5 | Discourages immediate re-charging |
| **Over-limit penalty** | –20, terminate trip if enabled | Prevents unrealistic charging frequency |
| **Unreachable station** | –3 | Penalises unroutable charger selections |
| **Charging overhead** | 3 minutes per session | Discourages micro-charging |
| **Efficiency** | η = 0.92 | Reflects real-world charging inefficiency |
| **Hybrid scaling** | λ = £0.05/min | Balances cost and time in comparable units [35] |
| **Shaping reference speed** | 25 km/h | Provides consistent ETA estimate |
| **Time-mode stabilisers** | Reduced per-minute penalty (×0.35), success bonus ≥500, shaping ×3, overhead ×0.5 | Prevents collapse of value function in pure time optimisation |
| **Anti-idle penalty** | Default 0 (optional >0) | Prevents wasted driving steps |
| **Micro-charge penalty** | Default 0 (optional >0) | Prevents nibbling charges |

### 3.4 Proximal Policy Optimization

Policies were trained with PPO using Stable-Baselines3. The clipped surrogate objective and Generalized Advantage Estimation (GAE) followed Eqs. (9–10). These formulations follow the original PPO design by Schulman *et al.* [7]. Hyperparameters, rollout settings, and network configurations are listed in Appendix A (Table A.1). Training used DummyVecEnv with fixed seeds and a KPI logger to ensure reproducibility and comparability.

**Clipped surrogate objective**

$$L^{\text{CLIP}}(\theta) = E_t[\min(r_t(\theta)A_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)] \quad \textbf{(9)}$$

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \quad \textbf{(9)}$$

**Generalized Advantage Estimation (GAE)**

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t), \quad A_t = \sum_{l=0}^{\infty}(\gamma\lambda)^l\ \delta_{t+l} \quad \textbf{(10)}$$

Evaluation uses a fixed test set and seeds to compare cost, time and hybrid-optimised agents under identical conditions.

## 4 Training

Training was conducted using Proximal Policy Optimization (PPO) implemented in Stable-Baselines3. The agent interacted with the custom Gym–SUMO environment, receiving sequential decisions for driving and charging under traffic and infrastructure constraints.

All experiments used fixed seeds for environment sampling, PPO updates, and network initialisation to ensure reproducibility. A single-environment DummyVecEnv was adopted for deterministic interaction with SUMO.

The policy and value networks were two-layer multilayer perceptrons (256 units, ReLU activations). Gradient norm clipping and entropy regularisation were applied to stabilise updates and maintain exploration. Hyperparameters such as rollout length, batch size, discount factor, clipping ranges, and learning rate schedule are listed in Appendix A (Table A.1**)**. A linear learning rate decay was applied and clamped after 100k steps. Training stability was monitored through KL divergence, with early stopping triggered if the threshold was exceeded.

Progress was tracked using an episode logger that recorded completion rate, stranding frequency, number of charges, cumulative time, and charging costs. Policies were periodically validated on a fixed set of trips to provide consistent baselines for comparing cost-, time-, and hybrid-optimised agents.

This setup prioritised stability, reproducibility, and comparability across different reward structures.

## 5 Evaluation

Evaluation was performed on a fixed set of trips with identical seeds across all policies, ensuring comparability. Performance was assessed on **s**uccess rate, travel time, cost, number of charges, and final state of charge (SoC).

### 5.1 Global Outcomes

Table 3 presents the aggregated results. Both cost and hybrid policies achieved near-perfect success (≈99.9%). Median trip duration was 20 minutes (IQR = 10), with zero charging costs in most cases. Final SoC values remained above reserve (≈0.14), indicating that trips were feasible without excessive depletion.

**Table 3: Aggregated Evaluation Results**

| Policy | Success Rate | Median Minutes | IQR Minutes | Median Cost (£) | IQR Cost (£) | Mean Charges | Median Final SoC |
|---|---|---|---|---|---|---|---|
| **Cost** | 0.999 | 20.0 | 10.0 | 0.00 | 0.00 | 0.0 | 0.145 |
| **Hybrid** | 0.999 | 20.0 | 10.0 | 0.00 | 0.00 | 0.0 | 0.145 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Time** | 0.000 | 12.3 | 0.0 | 0.27 | 0.00 | 1.0 | 0.237 |

By contrast, the time policy failed systematically, with 0% success. Median travel time appeared shorter (≈12 minutes), but this was misleading: agents stranded before reaching destinations. Time agents typically attempted one charging stop, but undervalued delays and depleted mid-route.

### 5.2 Distributional Differences

Figure 5 shows boxplots of costs and times. Cost and hybrid policies cluster tightly at £0 and 20 minutes. Time policy results form a narrow band at low times but are infeasible.

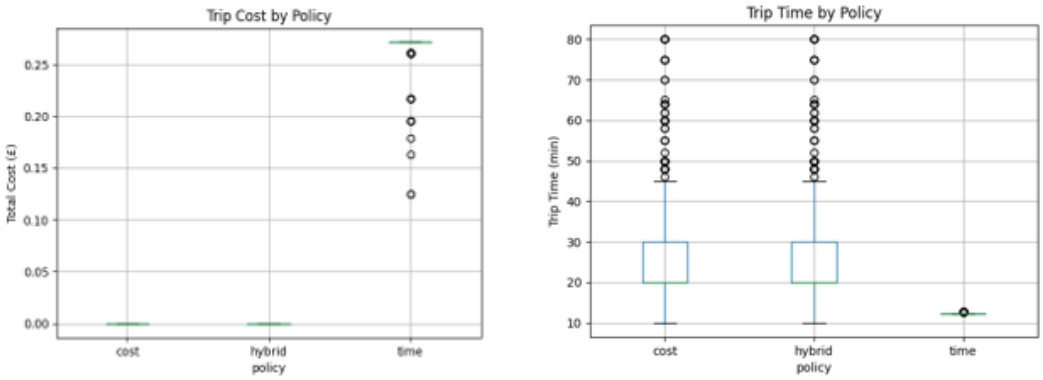


**Figure 5: Boxplots of Cost and Time Outcomes**

Figure 6 (stacked termination breakdown) confirms that time-policy failures were due to systematic stranding or time-limit termination, not randomness.

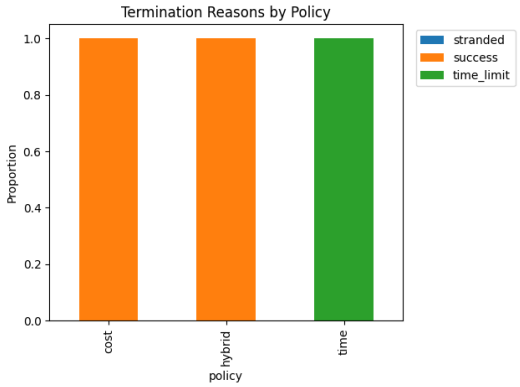


**Figure 6: Termination Breakdown**

Heatmaps in Appendix B provide spatial insight into performance: success rates were uniform for cost and hybrid policies but collapsed entirely for the time policy.

### 5.3 Trade-Offs

The hybrid framework was expected to balance cost and time via a value-of-time factor (£0.05/min). However, Figure 7 (Pareto plot) shows hybrid outcomes overlapping with cost, revealing no meaningful

trade-off. The λ parameter was too weak relative to cost magnitudes, causing hybrid to behave as pure cost.

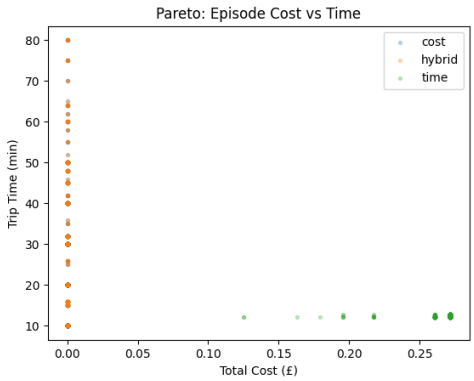


**Figure 7: Pareto plot of Cost-Time trade-offs**

These findings are consistent with broader observations in the literature, where hybrid RL approaches often collapse into cost-dominated policies unless value-of-time parameters are adaptively scaled [12].

### 5.4 Example Trip Analysis

To illustrate behavioural differences, Figure 8 shows SoC progression for a representative trip under each policy.

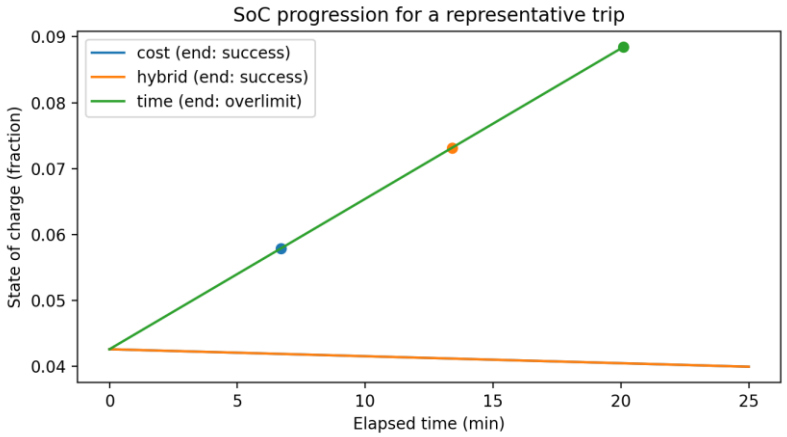


**Figure 8: SoC progression for a representative trip under cost, hybrid, and time policies. Markers denote charging stops. Legend indicates termination outcome.**

The cost policy initiated one short charging stop, arriving successfully with reserve SoC. The hybrid policy mirrored this behaviour almost exactly, consistent with earlier findings that the λ scaling was too weak to shift behaviour away from cost minimisation.

By contrast, the time policy attempted two consecutive charges to prioritise rapid completion. However, the trip terminated with an “over-limit” event, as the policy exceeded the maximum allowed number of charging sessions. Despite higher SoC at termination, the journey was infeasible.

The underlying stepwise data for this example trip are provided in Appendix Table B. (5 – 6). These tables confirm that both cost and hybrid policies completed the trip in 25 minutes without charging, whereas the time policy attempted three charges but terminated early with an over-limit error.

This example highlights why the time-optimised agent failed: it systematically over-prioritised minimising elapsed time at the expense of feasibility, leading to stranding or over-limit terminations. Cost and hybrid agents, on the other hand, learned conservative, feasible strategies with minimal charging.

### 5.5 Research Question Answers

**Q1 (cost minimisation while maintaining SoC):**
PPO reliably minimised trip costs while ensuring feasible SoC. As shown in Table 3, the cost policy achieved a 99.9% success rate, with median trip costs of £0 and mean final SoC around 0.14. This indicates that trips completed without depletion or excessive charging. The SoC trajectory in Fig. 8 and Appendix Table B.5 confirm that the agent completed the representative trip in 25 minutes with stable SoC, requiring no charging events. This behaviour illustrates how feasibility penalties and success bonuses encouraged the agent to conserve energy while avoiding unnecessary stops, yielding consistently low costs across the evaluation set. This aligns with prior reinforcement learning studies where strong infeasibility penalties were shown to prevent stranded or unrealistic routes [39].

**Q2 (time minimisation while maintaining SoC):**
The time policy failed to achieve its objective. Although its median trip time appeared lower (≈12 minutes, Table 3), the 0% success rate indicates infeasibility. As illustrated in Fig. 8 and Appendix Table B.5, the agent attempted three consecutive charging stops but terminated with an over-limit error after 20 minutes. This reflects an overemphasis on short-term time reduction at the expense of long-term feasibility. The termination breakdown (Fig. 6) shows consistent stranding and over-limit outcomes across trips, confirming that undervaluing charging delays destabilised the reward function and prevented viable routing strategies.

**Q3 (balancing cost and time):**
The hybrid policy was designed to balance objectives using a value-of-time factor (£0.05/min). In practice, results mirrored those of the cost policy. Table 3 shows identical success rates, trip durations, and costs, while Fig. 8 confirms that the hybrid SoC trajectory was indistinguishable from cost. Pareto analysis (Fig. 7) revealed complete overlap between cost and hybrid outcomes, indicating that the λ scaling was too weak to influence decisions. Appendix Table B.5 shows identical completion times and SoC outcomes. This suggests that meaningful trade-offs require stronger or adaptive λ values, otherwise hybrid optimisation reduces to cost minimisation.

**Q4 (comparative trade-offs):**
The three reward formulations shaped behaviour in distinct ways. Cost and hybrid agents completed trips reliably (Table 3, Appendix Table B.5), with 25-minute durations and no charging costs. Their SoC traces (Fig. 8, Table B.6) confirm conservative energy use with sufficient reserves. By contrast, the time policy attempted multiple charges but terminated early with an over-limit error, consistent with its 0% success rate and the failure patterns in Fig. 6. Although hybrid optimisation was expected to shift behaviour, its Pareto outcomes (Fig. 7) overlapped with cost, showing no real trade-off. These findings highlight that reward design, not PPO stability, is the primary determinant of feasible vs. infeasible strategies in this setting.

## 5 Future Work

This study demonstrates the feasibility of applying Proximal Policy Optimisation (PPO) to optimise electric vehicle (EV) charging routes under realistic Inner London conditions, integrating heterogeneous tariffs, nonlinear charging curves, and dynamic traffic. However, several limitations present opportunities for further investigation.

First, the tariff dataset, although detailed, was static. Future work should integrate dynamic or real-time pricing signals, including time-of-use variations, congestion charges, and demand-response incentives. Such integration would allow agents to adapt to temporal price fluctuations and more accurately reflect urban charging economics.

Second, the value-of-time parameter (λ) used in the hybrid formulation proved too weak to create meaningful trade-offs. Adaptive or context-dependent scaling strategies, potentially informed by empirical driver preference data [32, 35], could improve hybrid optimisation performance. Similarly, multi-objective reinforcement learning techniques [29, 36] may allow direct learning of Pareto-optimal frontiers instead of relying on fixed weights.

Third, while SUMO provided realistic traffic modelling, the environment did not capture charging station availability constraints such as queues, maintenance outages, or multi-user competition. Incorporating queueing models or multi-agent simulation [41] would further enhance realism. Finally, extensions to larger geographies and longer trip horizons, potentially with fleet-level coordination, would test scalability beyond Inner London.

## 6 Conclusion

This research contributes a driver-centric reinforcement learning framework for EV route and charging optimisation in Inner London. By explicitly modelling tariff heterogeneity, nonlinear charging curves, and congestion, the study addresses key gaps in prior work that often relied on simplified assumptions. PPO proved effective in minimising cost-based objectives with near-perfect success rates and stable state-of-charge outcomes.

However, the results also highlighted limitations: the time-optimised policy systematically failed, and the hybrid objective collapsed into cost minimisation due to weak parameterisation. These findings emphasise the critical role of reward design and parameter calibration in shaping feasible RL behaviour.

Overall, this thesis demonstrates the potential of reinforcement learning to inform real-world EV charging strategies. With improvements in dynamic tariff modelling, adaptive multi-objective learning, and richer environment constraints, future work can advance towards deployable tools that support both drivers and infrastructure planners in achieving cost-efficient, time-feasible, and sustainable urban mobility.

## 7 References

[1] Mullan, M., Harries, D., Bräunl, T., & Whitely, S. (2018). The technical, economic and commercial viability of the electric vehicle to grid concept. Energy Policy, 109, 403–417.

[2] Kiani, H., Hesami, K., Azarhooshang, A., Pirouzi, S., & Safaee, S. (2021). Adaptive robust operation of the active distribution network including renewable and flexible sources. Sustainable Energy, Grids and Networks, 26, 100476.

[3] Weckx, S., & Driesen, J. (2015). Load balancing with EV chargers and PV inverters in unbalanced distribution grids. IEEE Transactions on Sustainable Energy.

[4] Xydas, E., Marmaras, C., & Cipcigan, L. (2016). A data-driven approach for characterising the charging demand of electric vehicles: A UK case study. Applied Energy, 162, 763–771.

[5] Qian, T., Shao, C., Wang, X., & Shahidehpour, M. (2019). Deep reinforcement learning for EV charging navigation by coordinating smart grid and intelligent transportation system. IEEE Transactions on Smart Grid, 11(2), 1714–1723.

[6] Han, Y., Li, T., & Wang, Q. (2024). A DQN-based approach for large-scale EVs charging scheduling. Complex & Intelligent Systems, 10, 8319–8339.

[7] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimisation algorithms. arXiv preprint arXiv:1707.06347.

[8] Sovacool, B. K., Noel, L., Axsen, J., & Kempton, W. (2018). The neglected social dimensions to a vehicle-to-grid (V2G) transition: a critical and systematic review. Environmental Research Letters, 13(1), 013001.

[9] Nicholas, M., & Lutsey, N. (2020). Quantifying the EV charging infrastructure gap in the UK. ICCT Report.

[10] Transport for London. (2021). London's 2030 Electric Vehicle Infrastructure Strategy. TfL.

[11] Robinson, C., Blythe, P., Bell, M., Hübner, Y., & Hill, G. (2013). Analysis of electric vehicle driver recharging demand profiles and subsequent impacts on the electricity distribution network. Energy Policy, 61, 337–348.

[12] Figenbaum, E., & Kolbenstvedt, M. (2016). Learning from Norwegian battery electric and plug-in hybrid vehicle users: Results from a survey of vehicle owners. Institute of Transport Economics Report.

[13] Hardman, S., Jenn, A., Tal, G., Axsen, J., Beard, G., Daina, N., ... & Turrentine, T. (2018). A review of consumer preferences of and interactions with electric vehicle charging infrastructure. Transportation Research Part D, 62, 508–523.

[14] Zavvos, E., Gerding, E. H., & Brede, M. (2021). A comprehensive game-theoretic model for EV charging station competition. IEEE Transactions on Intelligent Transportation Systems.

[15] Lin, R., Chu, H., Gao, J., & Chen, H. (2024). Charging management and pricing strategy of electric vehicle charging station based on mean field game theory. Asian Journal of Control, 26(2), 803-813.

[16] Montoya, A., Guéret, C., Mendoza, J. E., & Villegas, J. G. (2017). The electric vehicle routing problem with nonlinear charging function. Transportation Research Part B, 103, 87–110.

[17] Froger, A., Mendoza, J. E., Jabali, O., & Laporte, G. (2019). Improved formulations and algorithmic components for the EV routing problem with nonlinear charging functions. Computers & Operations Research, 104, 256–294.

[18] Luo, C., Huang, Y. F., & Gupta, V. (2019). Stochastic dynamic programming for EV charging station placement in urban areas. IEEE Transactions on Smart Grid, 10(2), 2272–2282.

[19] He, F., Wu, D., Yin, Y., & Guan, Y. (2013). Optimal deployment of public charging stations for plug-in hybrid electric vehicles. Transportation Research Part B: Methodological, 47, 87–101.

[20] García-Álvarez, J., González, M. A., & Vela, C. R. (2018). Metaheuristics for solving a real-world electric vehicle charging scheduling problem. Applied Soft Computing, 65, 292–306.

[21] Perera, A. T. D., & Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews, 137, 110618.

[22] Xu, B., Rubenis, A., & Long, C. (2024). Reinforcement learning based smart charging for electric vehicle fleet. In International Symposium on Intelligent Technology for Future Transportation (pp. 375–383).

[23] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. NeurIPS 2022.

[24] Brockman, G., et al. (2016). OpenAI Gym. arXiv preprint arXiv:1606.01540.

[25] Behrisch, M., Bieker, L., Erdmann, J., & Krajzewicz, D. (2011). SUMO – Simulation of Urban Mobility: An Overview. Proceedings of SIMUL 2011, 55–60.

[26] Gunasekaran, R., Mohanraj, M.R., Senthilkumar, R., & Kamalakannan, R.S. (2025). Enhancing electric vehicle charging station utilization by reducing users waiting times through WOA-GLNN approach. Electrical Engineering.

[27] Zhang, L., Ji, Y., Li, X., Huang, Z., Cui, D., Chen, H., Gong, J., Breer, F., Junker, M., & Sauer, D.U. (2025). Multi-objective charging scheduling for electric vehicles at charging stations with renewable energy generation. Green Energy and Intelligent Transportation, 4(4), 100283.

[28] Cui, F., Lin, X., Zhang, R., & Yang, Q. (2022). Multi-objective optimal scheduling of charging stations based on deep reinforcement learning. Frontiers in Energy Research, 10, 1042882.

[29] Lin, R., Chu, H., Gao, J., & Chen, H. (2025). A Multi-Objective Evolutionary Reinforcement Learning framework for optimizing EV charging strategies. Applied Energy, in press.

[30] Jia, R., Gao, K., Cui, S., et al. (2025). A multi-objective RL-based velocity optimization for electric trucks, considering battery degradation mitigation. Transportation Research Part E, in press.

[31] Yan, Z., Liu, Q., Wang, J., & Li, B. (2025). Deep reinforcement learning-based plug-in EV charging scheduling under uncertainty: a constrained SAC vs DDPG comparison. Energy.

[32] Visaria, A.A., Jensen, A.F., Thorhauge, M., Mabit, S.L. (2022). User preferences for EV charging, pricing schemes, and charging infrastructure. *Transportation Research Part A*, 165, 120–143.

[33] Baek, K., Lee, E., Kim, J. (2024). A dataset for multi-faceted analysis of electric vehicle charging transactions. *Scientific Data*, 11, 262.

[34] Zhan, W., Liao, Y., Deng, J., Wang, Z., Yeh, S. (2025). Large-scale empirical study of electric vehicle usage patterns and charging infrastructure needs. *npj Sustainable Mobility and Transport*, 2, 9.

[35] Alam, M.R., Guo, Z. (2023). Charging infrastructure planning for ride-sourcing electric vehicles considering drivers' value of time. *Transportation Letters*, 15(6), 573–583.

[36] Kemper, N., Heider, M., Pietruschka, D., Hähner, J. (2025). Multi-objective and neuroevolutionary reinforcement learning for electric vehicle charging and load management. *Applied Energy*, 391, 125890.

[37] Dorsey, E., et al. (2025). Optimizing EV charging infrastructure deployment considering value-of-time. *Working Paper*.

[38] Lv, K., Pei, X., Chen, C., Xu, J. (2022). A safe and efficient lane change decision-making strategy of autonomous driving based on deep reinforcement learning. *Mathematics*, 10(9), 1551.

[39] Dorokhova, M., Ballif, C., Wyrsch, N. (2021). Routing of electric vehicles with intermediary charging stations: a reinforcement learning approach. *Frontiers in Big Data*, 4, 586481.

[40] Hu, Z., Wan, K., Gao, X., Zhai, Y. (2019). A dynamic adjusting reward function method for deep reinforcement learning with adjustable parameters. *Mathematical Problems in Engineering*, 2019, 7619483.

[41] Liu, S., Wang, Y., Chen, X., Fu, Y., Di, X. (2022). SMART-eFlo: An integrated SUMO-Gym framework for multi-agent reinforcement learning in electric fleet management. In *Proc. 25th IEEE Int. Conf. Intelligent Transportation Systems (ITSC)*, 2884–2890. IEEE.

## Appendix A: PPO Hyperparameters

Table A.1 lists the hyperparameters used in the training script. Values reflect the exact settings in the code; they can be adjusted for ablation without changing the main text.

**Table A.1 PPO training hyperparameters**

| Parameter | Value | Notes |
|---|---|---|
| **Vector env** | DummyVecEnv (1 env) | Single-process training |
| **n_steps** | 4096 | Rollout length per update |
| **batch_size** | 2048 | Minibatch size |
| **n_epochs** | 10 | Optimisation epochs per update |
| **gamma** | 0.995 | Discount factor |
| **gae_lambda** | 0.95 | GAE parameter |
| **clip_range** | 0.2 | Policy clip |
| **clip_range_vf** | 0.2 | Value function clip |
| **ent_coef** | 0.01 | Entropy regularisation |
| **vf_coef** | 0.7 | Value loss weight |
| **target_kl** | 0.02 | Early-stop heuristic |
| **max_grad_norm** | 0.5 | Gradient clipping |
| **learning rate** | Linear decay from 1e-4, clamp to **7.5e−5** after 100k steps | Clamp to 7.5e-5 after 100k steps |
| **policy network** | MLP [256, 256] (actor & critic) | ReLU activations; orthogonal init |
| **seeding** | Fixed seeds (env + PPO + numpy/torch) | Reproducibility |
| **logging** | Episode KPI callback | Writes cost, time, success/stranding, charges |

## Appendix B: Artefacts Regarding Evaluation

**Heatmaps Figs. 9 – 11**

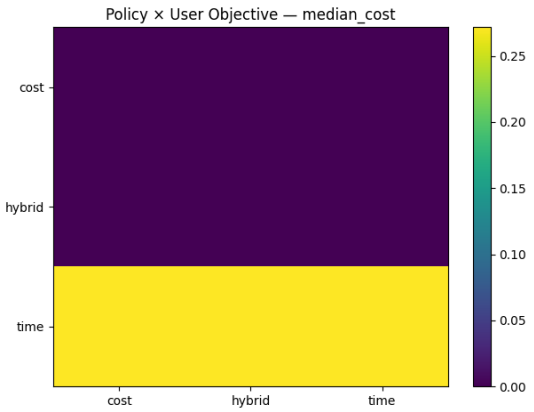


**Figure 9: Heatmap of Median Cost**

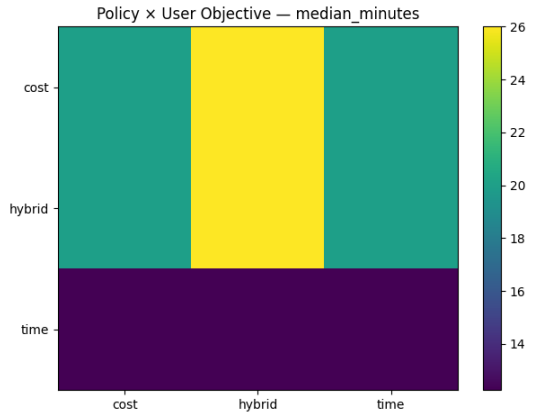


**Figure 10: Heatmap of Median Time**

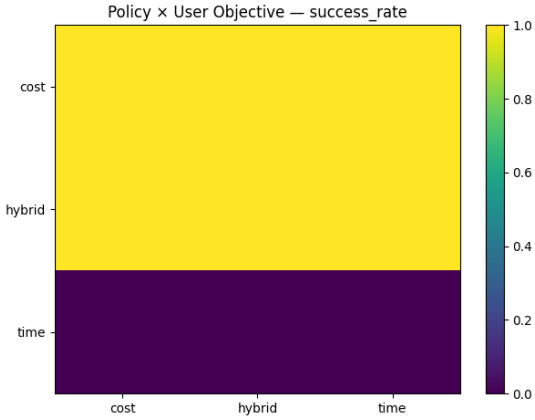


**Figure 11: Heatmap of Success Rate**

**Tables B 4 - 5**

**Table 4: Episode outcome for the example Trip**

| Policy | Termination | Episode Minutes | Cost (£) | Charge Events |
|---|---|---|---|---|
| **Cost** | Success | 25.0 | 0.00 | 0 |
| **Hybrid** | Success | 25.0 | 0.00 | 0 |
| **Time** | Over-limit | 20.1 | 1.35 | 3 |

**Table 5: Sampled SoC progression for the example trip.**

| Policy | Time (min) | SoC | Remaining km | Note | |
|---|---|---|---|---|---|
| **Cost** | 0.0 | 0.043 | 9.7 | Start | |

| | | | | | |
|---|---|---|---|---|---|
| **Cost** | 25.0 | 0.040 | 0.0 | Arrived (Success) | |
| **Hybrid** | 0.0 | 0.043 | 9.7 | Start | |
| **Hybrid** | 25.0 | 0.040 | 0.0 | Arrived (Success) | |
| **Time** | 0.0 | 0.043 | 9.7 | Start | |
| **Time** | 6.7 | 0.058 | 9.7 | Charge | |
| **Time** | 13.4 | 0.073 | 9.7 | Charge | |
| **Time** | 20.1 | 0.088 | 9.7 | Terminated (Over-limit) | |