Smart Transportation Without Neurons - Fair Metro Network Expansion with Tabular Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

We address the Metro Network Expansion Problem (MNEP), a subset of the Transport Network Design Problem (TNDP), which focuses on expanding metro systems to satisfy travel demand. Traditional methods have relied on exact and heuristic approaches that require expert-defined constraints to reduce the search space and enable tractability. Recently, reinforcement learning (RL), particularly deep reinforcement learning (Deep RL), has emerged as a powerful alternative due to its effectiveness in optimizing complex sequential decision-making processes. However, Deep RL methods can be computationally expensive, environmentally costly and hard to interpret. In this paper we re-formulate the MNEP as a Markov Decision Process (MDP), and solve it through tabular Q-Learning. By using a redefined MDP and a tabular RL approach, we achieve similar performance to Deep RL, with substantially fewer training episodes, offering the added benefit of greater interpretability. Furthermore, we incorporate diverse social equity criteria into the reward functions, balancing efficiency with fairness, thus highlighting the versatility of our method. Our approach is evaluated in real-world settings—specifically in Xi'an and Amsterdam—where it demonstrates competitive results, reducing the total training episodes by a factor of 18 and total carbon emissions by a factor of 12 on average. Our approach provides a replicable, interpretable, and resource-efficient solution, with potential applicability to other combinatorial optimization problems.

1 Introduction

Public transport is fundamental to modern, fast-paced lifestyles, as it enables citizens to participate in employment, education, healthcare, and social activities Martens (2016). However, planning public transport networks is especially challenging due to physical, social, economic and legal constraints that complicate the creation of new transport routes, or the expansion of existing ones. Additionally, sustainability and equity are values that increasingly shape the design of public transport networks. Modern transportation systems must be accessible, ensuring that citizens of all locations, socioeconomic statuses, and ages can benefit from these services Martens (2016). They also need to be efficient, as they must cover actual demand for mobility rather than being designed arbitrarily. Efficiency is also vital for sustainability: buses with low passenger loads can have a higher environmental impact per passenger than cars Lowe et al. (2009), and low ridership can degrade the quality of transit systems over time Mohring (1972). These trade-offs add further complexity to transport design problems, leading to the need for increasingly sophisticated solutions.

The Transport Network Design Problem (TNDP) is an NP-hard combinatorial optimization problem that addresses the design of public transport, by maximizing maximize total travel demand satisfaction Farahani et al. (2013). For metro systems, a specific subset of TNDP, known as the Metro Network Expansion Problem (MNEP), is central to expanding existing metro lines within cities Wei et al. (2020); Wang et al. (2023); Su et al. (2024). Metro networks are especially important in modern cities for their speed, reliability, and high passenger capacity compared to other traditional modes of public transport Wang et al. (2023).

MNEP are a subset of problems of TNDP, focused on expanding an existing metro network in a city. Metro lines generally cover long distances, cross multiple urban zones, and are typically designed as relatively straight routes without excessive meandering Wei et al. (2020). As a distinct sub-problem within TNDP, MNEP introduces additional constraints specific to metro network design.

Traditionally, TNDP problems have been approached with integer optimization and heuristic algorithms Laporte & Pascoal (2015); Owais & Osman (2018), which require extensive expert-defined constraints to reduce the search space for tractability. Recently, the Metro Network Expansion Problem (MNEP) has been framed as a sequential decision-making problem, leveraging Reinforcement Learning (RL) to derive optimal solutions Wei et al. (2020). RL is well-suited for sequential decision-making with multiple objectives, such as efficiency and fairness, and has been successfully applied to combinatorial optimization problems Darwish et al. (2020); Raman et al. (2021); Jullien et al. (2022). Unlike traditional methods, RL can explore the search space flexibly by optimizing a reward function, avoiding the need for exponentially increasing constraints.

Given the large state-action spaces in many problems, the complexity of Reinforcement Learning (RL) may seem justified. Recently, Deep Reinforcement Learning (Deep RL) has shown promise in scaling combinatorial optimization, learning policy representations that autonomously identify key features and achieving state-of-the-art results in real-world problems Mazyavkina et al. (2021); Neustroev et al. (2022); Xu et al. (2022).

While advances in computing power and algorithmic research suggest that RL could transform problems like MNEP, we argue that Deep RL is not always the ideal solution. Its substantial training time and environmental costs are becoming increasingly significant with the widespread deployment of AI systems Anthony et al. (2020); Strubell et al. (2020); Patterson et al. (2021); Krishnan et al. (2022). Given that MNEP and similar problems are static optimizations with limited features, complex neural network structures may not be necessary for effective policy training. This is supported by findings in other machine learning domains Cuccu et al. (2019).

In this paper, we propose that traditional, tabular-based methods in reinforcement learning can effectively address complex problems like MNEP, provided the problem is framed appropriately. We show that a tabular approach can achieve competitive performance compared to deep-learning methods, reducing training time significantly in two real-world environments (Xi'an and Amsterdam), while also offering greater interpretability than black-box deep learning models.

To further demonstrate the potential of tabular RL, we examine various aspects of social equity in MNEP by employing a range of reward functions based on diverse concepts of social good. We achieve this by expanding the state-of-the-art RL formulation of the MNEP, to incorporate fairness criteria aligned with diverse definitions of social welfare.

In summary, we make the following contributions:

- We reformulate the Transport Network Design and Metro Network Expansion problems as Markov Decision Processes with two-stage actions, significantly reducing the action space compared to existing approaches.
- We bridge machine learning and transport planning research by extending the RL framework to integrate considerations of social good, balancing efficiency and fairness.
- We propose a Monte Carlo Tabular Reinforcement Learning algorithm for MNEP, designed to require fewer training episodes than deep learning models.
- We validate our method in two real-world settings—Xi'an, China, and Amsterdam, Netherlands—demonstrating comparable performance to state-of-the-art Deep RL methods, with an 18-fold reduction in training episodes and a 12-fold reduction in CO_2 emissions.
- We provide all code, datasets, and hyperparameter settings to replicate our results and enable application to other combinatorial optimization problems[1].

The remainder of the paper is structured as follows: First, we position our work in the context of previous research (Section 2) and re-formulate the MDP formulation of the MNEP (Section 3). We continue by describing the tabular model and the proposed social-welfare reward functions (Section 4) and the real-world environments used in our experiments (Section 5). Finally, we present and discuss our results (Section 6).

[1]Github: https://github.com/*****/ (retracted for anonymous submission)

2 Related Work

We outline previous work on the TNDP, reinforcement learning for combinatorial optimization, and the analysis of fairness in transportation.

2.1 Transport Network Design Problem

Traditionally, the Transport Network Design Problem (TNDP) has been approached through a combination of integer optimization techniques and heuristic methods, including the use of pre-defined or dynamically discovered corridors Laporte & Pascoal (2015); Zarrinmehr et al. (2016); Gutiérrez-Jarpa et al. (2018), simulated annealing Fan & Machemehl (2006); Ahern et al. (2022), bee colony optimization Yang et al. (2007); Szeto & Jiang (2014), and genetic algorithms Owais & Osman (2018); Nayeem et al. (2018).

While these approaches have produced promising results in early studies, they have notable limitations. To make the problem tractable for solvers, they often restrict the search space by either enforcing a long list of environment-specific constraints or by setting, or searching for, a predefined set of corridors. This restriction provides obstacles in application in large, real-world urban environments with diverse characteristics. More critically, narrowing the search space in this manner can exclude high-quality solutions that lie outside of these constraints.

2.2 Reinforcement Learning for Transport Network Design

Reinforcement Learning (RL) has proven effective for making optimal long-term sequential decisions. Through straightforward reward mechanisms, an agent learns to understand its impact on the environment via trial-and-error, making RL well-suited for tackling real-world NP-hard combinatorial optimization tasks by leveraging demonstration and experience, without the need for expert prior knowledge Mazyavkina et al. (2021); Wang & Tang (2021); Bengio et al. (2021); Jullien et al. (2022). Although combinatorial optimization problems can also be approached with Supervised Learning (SL), recent studies have shown that RL can generalize more effectively than SL in common problems such as the Travelling Salesman Problem Bello et al. (2017); Deudon et al. (2018) and Vehicle Routing Nazari et al. (2018); Kool et al. (2018).

Despite RL's growing utility in combinatorial optimization, its use in transport network design has been limited. Darwish et al. used a policy gradient method to design bus lines, exploring the Pareto front between customer satisfaction and operational costs Darwish et al. (2020). Wei et al. employed a pointer-based model to solve the Transit Network Design Problem (TNDP), achieving superior performance in meeting demand satisfaction Wei et al. (2020). Additionally, Multi-objective Reinforcement Learning has been applied to the TNDP to balance efficiency with accessibility Zhang et al. (2024); Michailidis et al. (2023).

Most work on the Transit Network Design Problem (TNDP) and the closely related Metro Network Expansion Problem (MNEP) has focused on complex deep reinforcement learning (Deep RL) models. This paper, however, challenges the necessity of such black-box models for problems where interpretability is crucial for decision-makers. We reformulate the Markov Decision Process to significantly reduce the action space without restricting the solution space, enabling a simpler, Monte Carlo-based tabular reinforcement learning approach. Our method is then benchmarked against the state-of-the-art Deep RL approach for MNEP Wei et al. (2020).

2.3 Social Equity in Transport Network Design

Adopting notions of social equity in transport network design is challenging to optimize due to its multi-dimensional nature Behbahani et al. (2019) and the inherent moral judgments involved van Wee (2011). Drawing on prior research in urban transportation, we identify three key decisions necessary to incorporate fairness: utility measure, dimension, and fairness theory.

Utility measure: This is commonly achieved by establishing accessibility metrics, such as the number of reachable opportunities Pereira et al. (2019); van der Veen et al. (2020); Hernandez (2018), the affordability of accessing them Farber et al. (2014), or a combination of both El-Geneidy et al. (2016).

Dimension: Fairness can be assessed along spatial dimensions, where disparities are evaluated across different geographic or administrative units Pereira et al. (2019); Delmelle & Casas (2012), or through group-based measures, where groups are defined by socio-economic characteristics (e.g., income, race) van der Veen et al. (2020); Pyrialakou et al. (2016); Cheng et al. (2021).

Fairness theory: Multiple theories of fairness and equity inform transport network design Behbahani et al. (2019). Most approaches fall under horizontal fairness—aiming for equal utility across all units or groups—or vertical fairness, which prioritizes groups or areas in greater need van Wee (2011).

Despite these theoretical analyses, comprehensive application of fairness frameworks within machine learning for TNDP remains limited. Nonetheless, prior work has made initial attempts to integrate equity considerations. For example, Ramachandran et al. explore the efficiency-equity trade-off in graph augmentation using RL, applying their approach to Chicago's transportation network Ramachandran et al. (2021). Tedjopurnomo et al. compare bus line designs for advantaged and disadvantaged groups, though not using RL Tedjopurnomo et al. (2022). Wei et al. account for equity by designing a weighted reward that balances travel demand with an area's development index, though this measure is implemented within the reward function and analyzed only minimally for its impact. The same approach is used by Zhang et al., who add one more component to the reward function.

Our paper presents the first attempt to bridge the gap between transportation fairness research and RL-based transport network design in a comprehensive framework. We design fairness-based rewards based on Behbahani et al. definition, which targets an equitable distribution of benefits introduced by new transport lines. This framework is adaptable to various utility measures; in this study, we focus on Origin-Destination flows due to their relevance for mobility demand, rather than accessibility. Our analysis is done on a socio-economic group dimension, and we provide diverse reward functions that cover different fairness notions.

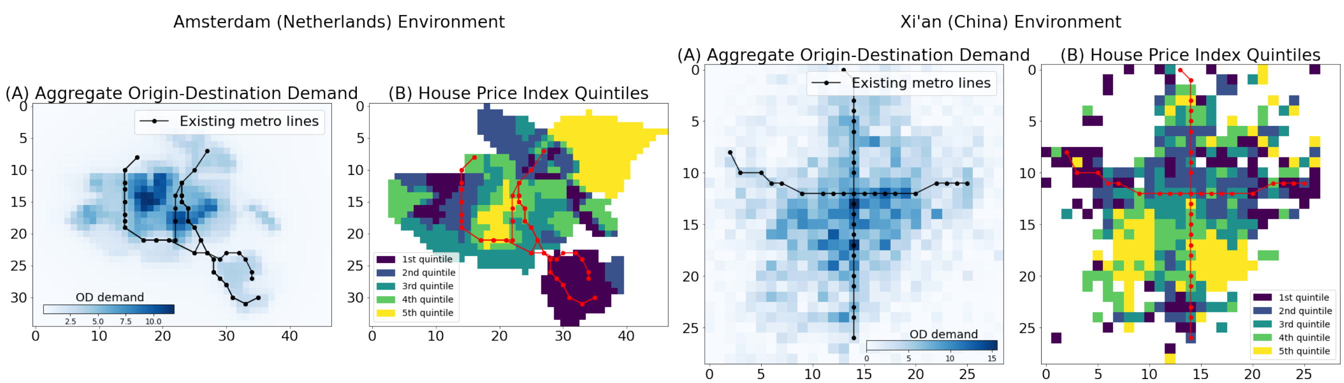


Figure 1: Two real-world case studies where the Metro Network Expansion Problem (MNEP) can be applied. The left side features Amsterdam, Netherlands, with each grid cell representing aggregate origin-destination demand (visualized using a blue colormap in panel A), along with the city's existing metro lines and housing price quintiles (panel B). On the right, similar data is displayed for Xi'an, China.

3 The Metro Network Expansion Problem

The Metro Network Expansion Problem (MNEP) is a subproblem of the Transport Network Design Problem (TNDP). Within the TNDP framework, the primary objective is to expand the transport network by constructing a new line that maximizes captured travel demand, considering unmet demand and connections to the existing network.

In traditional formulations of TNDP and MNEP, the city is modeled as a two-dimensional grid environment with n rows and m columns, $H^{n \times m}$. The aim is to identify a set of adjacent cells $Z = \{z_1, z_2, \ldots, z_T \mid z_i \in H, \forall i = 1, 2, \ldots, T\}$, which sequentially connect to form a new metro line, in order to maximize the total captured demand. This demand is represented by an Origin-Destination (OD) matrix, $OD^{|H| \times |H|}$ Guihaire & Hao (2008); Farahani et al. (2013). Here, $OD[i, j]$ denotes the travel demand from grid cell i to grid cell j.

In the MNEP, the OD matrix is assumed to be symmetric and deterministic, remaining constant throughout the optimization process.

The size of set Z is limited by a construction budget B , and a maximum number of stations T. We define a function $U(Z)$ that calculates the total added benefit of the generated line Z. In the traditional MNEP, $U(Z)$ is defined as the total sum of satisfied demand. The optimization problem is then defined as follows. Find the set of connected cells Z, such that:

$$\begin{aligned} \max \quad & U(Z) = \sum_i \sum_j OD[z_i, z_j], i \neq j \\ \text{s.t.} \quad & cost(Z) \leq B \\ & |Z| \leq T \end{aligned} \tag{1}$$

Here, the constraints B and T are strict, meaning that the new metro line must not exceed the specified budget or the total number of allowable stations.

The structural configuration of the metro line depends on the type of transport, which can be directed, as observed in bus or tram networks, or undirected, as is typical in metro or subway systems. The focus of our study is the design of metro networks, hence we tackle the Metro Network Expansion Problem (MNEP) Wei et al. (2020).

3.1 Social Equity in the Metro Network Expansion Problem

The traditional MNEP primarily seeks to maximize total demand coverage, often overlooking the equitable distribution of benefits across various communities within the city. Prior work on reinforcement learning (RL) in this context also tends to prioritize efficiency and adopt a predominantly *utilitarian* approach Wei et al. (2020). Here, we demonstrate that RL can effectively optimize for a wider array of objectives that encompass essential principles of social equity, as defined in transport planning literature. In addition to *utilitarianism* (Equation (1)), we emphasize two additional equity principles: *equal sharing of benefits* and *Rawlsian justice* as articulated by Rawls' theory of justice Behbahani et al. (2019).

Our focus centers on ensuring fairness in the allocation of satisfied Origin-Destination demand facilitated by the new line, paying particular attention to its distribution across different socioeconomic groups.

We first define a set of groups G, based on socioeconomic indicators such as income, development index, and education. Each cell $h \in H^{n \times m}$ in the environment is associated with a group $g \in G$. We adjust the objective function for each fairness notion accordingly, defining a utility function $U(Z, g)$ for each group $g \in G$, which returns the satisfied OD demand of line Z for group g.

Each cell $h \in H^{n \times m}$ of the environment is associated with a group $g \in G$. We adjust the objective function for each fairness notion accordingly. We define a utility function $U(Z, g), g \in G$, which returns the satisfied OD demand of line Z for group g.

Equal Sharing: This egalitarian objective aims to equalize the added benefits of the transport line among groups in a city, commonly referred to as horizontal equity. In theory, equal sharing is achieved by minimizing the absolute differences between group utilities:

$$\min \sum_i \sum_j |U(Z, g_i) - U(Z, g_j)|, g_i, g_j \in G, i \neq j \tag{2}$$

To implement fairness objectives in practice, we need to also incorporate total reward as, theoretically, Equation (2) could be minimized when all group utilities are 0. To address this, we encapsulate the equal-sharing notion using the Generalized Gini Index (GGI) Siddique et al. (2020).

$$U(Z) = GGI(Z, W) = \sum_i^{|G|} W_i, U(Z, \sigma(G)i), \tag{3}$$

where σ is a permutation that sorts the groups in G in descending order based on their utility prior to line creation, and W_i are non-increasing weights (i.e., $W_1 > W_2 > \cdots > W|G|$) normalized to sum to 1.

Rawls' Theory of Justice: This approach aims to maximize benefits for the most disadvantaged group.

$$\max(U(Z, g_{min})), \tag{4}$$

where g_{min} represents the most disadvantaged group within G. In this paper, we define groups based on a house-price index as a proxy for area development, with g_{min} as the group with the lowest house price index. Lower house price indexes are used as a proxy to identify the poorer areas of a city.

To apply this notion as a reward in RL, we set the reward function as $U(Z) = U(Z, g_{min})$. In Figure 1, we illustrate the real-world environments of Amsterdam and Xi'an where we apply the TNDP. We detail the environments in Section 5.

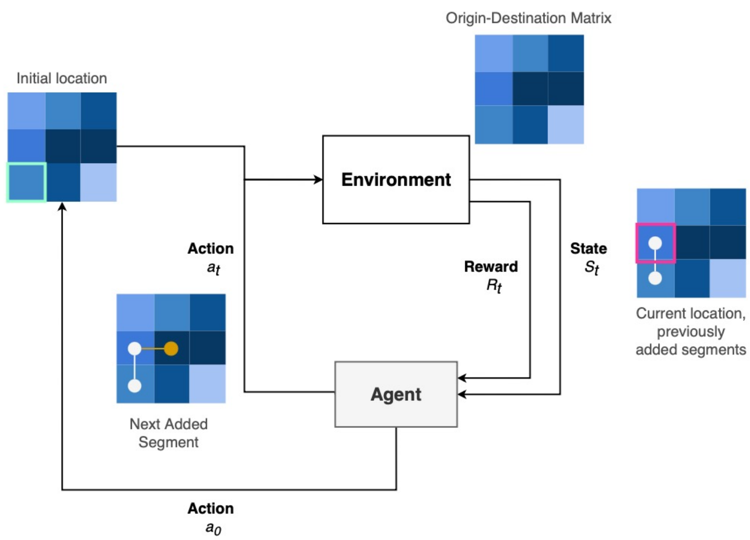


Figure 2: In the Metro Network Expansion Problem (MNEP), a reinforcement learning (RL) agent sequentially adds transport segments to the network. Each action represents the addition of a segment at a specific location, with rewards based on the demand met by that segment. The objective is to maximize the cumulative reward from all added segments.

4 Methods

We define the Metro Network Expansion Problem (MNEP) as a Markov Decision Process (Section 4.1) and describe the Tabular Q-Learning algorithm we use to solve it (Section 4.2).

4.1 Metro Network Expansion MDP

Recent approaches to the MNEP apply reinforcement learning (RL) by encoding each city grid cell as a potential action for the agent, resulting in a fixed action space of size $|H|$ at each timestep Wei et al. (2020); Su et al. (2024). This setup increases the search space as grid size grows. While physical constraints mask certain actions to limit selectable cells at each timestep, this masking occurs only after the forward pass, immediately before the softmax layer Wei et al. (2020); Su et al. (2024). As a result, the policy network must still process all potential cells in every state.

We argue that this complexity is unnecessary. Instead, we propose a two-stage solution: first, the agent selects a *starting cell*—the initial location for placing the first station on a metro line. The agent then navigates the grid by choosing among eight possible movement directions (north, south, east, west, and the four diagonal directions). Each movement forms a segment of the metro line, with the newly entered cell designated as the next station location.

This adjustment substantially reduces the action space to always be of size 8, for all timesteps except for the first one, regardless of the size of the grid. Furthermore, we simplify the state representation to the agent's current location, which can be efficiently encoded in a table with rows corresponding to the total number of cells in the city grid.

The MNEP is defined as an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu \rangle$, where $\mathcal{S}$ is the agent's current location, $\mathcal{A}$ represents the next movement direction, $\boldsymbol{\mathcal{R}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ indicates the demand satisfied by the last action. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a deterministic state transition functionand. Finally, $\mu : \mathcal{S} \to [0, 1]$ is a probability distribution over the starting state, which can be learned, kept constant, or randomized.

Given the discrete and episodic nature of the problem, we set the discount factor $\gamma = 1$, and the transition function $\mathcal{P}$ remains deterministic. Figure 2 illustrates this formulation.

The actions the agent can take at any timestep are further constrained by feasibility rules $F(Z_t)$, expressed by directional constraints, as in previous works Wei et al. (2020). These constraints include prohibitions on revisiting cells, preventing movement beyond grid boundaries, and restricting the agent to a singular directional movement, thereby avoiding cyclical paths (resembling a large-distance metro line). Detailed information on the feasibility rules can be found in Appendix A, the accompanying code and in prior literature Wei et al. (2020); Zhang et al. (2024).

4.2 Tabular Q-Learning for MNEP

We propose a tabular Q-learning algorithm for metro network expansion, using the MDP formulation outlined above. A single reinforcement learning (RL) agent (a) selects an initial cell and (b) extends the metro line by connecting adjacent cells in all possible directions. We apply a Monte Carlo-based method to iteratively update a Q-table through repeated environment interactions.

Selecting the Initial Cell At the start of each episode, the agent selects the initial state S_0 (starting point for the metro line) using an ϵ-greedy approach. In exploration, it picks a random cell; in exploitation, it selects the cell maximizing the expected return. The value of each cell as a starting point is represented by $Q_{s_0} \in \mathbb{R}^{|H|}$, giving one Q value for each grid cell. The exploration rate begins high ($\epsilon = 1$) and decays linearly until it reaches a minimum value ($\epsilon = 0.01$).

Action Selection and Transition At each timestep t, the agent selects an action A_t using ϵ-greedy selection based on values $Q \in \mathbb{R}^{|H| \times 8}$. Out-of-bound movements and other invalid actions are masked. After selecting an action A_t, the agent observes a reward R_t and deterministically transitions to a new state S'. This transition (S_t, A_t, R_t) is stored in an episodic list, tracking the agent's path, which is later used to perform Monte Carlo updates. Episodes end when one of three terminal conditions is met: (a) no available directions remain, (b) the budget is exhausted, or (c) the maximum number of allowed stations is reached.

Reward Calculation The reward at each timestep t reflects the additional demand met by the new metro segment, calculated in two steps. First, the direct demand between the new station and all previously existing stations on the line is computed. Next, if connections between the new metro line and existing lines are identified, the reward is increased by the additional transfer demand between each station of the existing line and each station of the extended line Wei et al. (2020). The total reward is the sum of these two components.

$$R_t = \underbrace{U(Z_t)}_{\text{direct demand}} + \underbrace{\sum_{l \in L} \mathbb{1}_{connect}(Z_t, l) \cdot U(l \times Z_t)}_{\text{transfer demand}}, \tag{5}$$

where $Z_t = z_1, ..., z_t$ is the set of all stations in the current line up to time t, L is the set of all existing metro lines, S_l is the set of stations in existing line l, $\mathbb{1}_{connect}(z_t, l)$ is an indicator function that equals 1 if station z_t connects with line l (shares a cell), and 0 otherwise.

Monte-Carlo Returns and Policy Update At the end of each episode, the agent updates the Q-values using Monte Carlo estimation. First, the total discounted return, denoted by J (we use J here to avoid confusion with the group set G, departing slightly from standard RL notation), is calculated. Using this return J, the agent then updates the Q-values accordingly. Finally, the Q-value for the initial state S_0 is updated separately, also based on the return J.

$$\begin{aligned} Q(S_t, A_t) &\leftarrow Q(S_t, A_t) + \alpha[J - Q(S_t, A_t)] \\ Q_{s_0}(S_0) &\leftarrow Q_{s_0}(S_0) + \alpha[J - Q_{s_0}(S_0)] \end{aligned} \tag{6}$$

In Algorithm 1 we show the pseudocode of the proposed method.

Algorithm 1 Tabular Metro Network Expansion with Monte-Carlo Updates

Parameters: B, T, α, γ ▷ Budget, total stations, RL parameters
Initialize $Q(s, a)$, Q_{s_0} for all cells s, actions a, empty $Episode$, $TotalCost \leftarrow 0$, and $ActionMask$ of ones.
for each episode **do**
  Select S_0 via ϵ-greedy from Q_{s_0}; add S_0 to Z
  for each step t **do**
    Choose A_t with ϵ-greedy, considering $ActionMask$
    Execute A, receive reward R, observe next state S'
    Append (S, A, R) to $Episode$, add z_t to Z, update $TotalCost$, $ActionMask$, $S \leftarrow S'$
    if $SUM(ActionMask) = 0$ OR $TotalCost \geq B$ OR $t \geq T$ **then break**
    end if
  end for
  Initialize $J \leftarrow 0$
  for each step (S_t, A_t, R_t) in $Episode$ from last to first **do**
    $J \leftarrow \gamma J + R_t$
    if (S_t, A_t) is first in $Episode$ **then**
      $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(J - Q(S_t, A_t))$
    end if
  end for
  Update $Q_{s_0}(S_0) \leftarrow \alpha(J - Q_{s_0}(S_0))$
  Reset $TotalCost$, $Episode$, Z, and $ActionMask$
end for

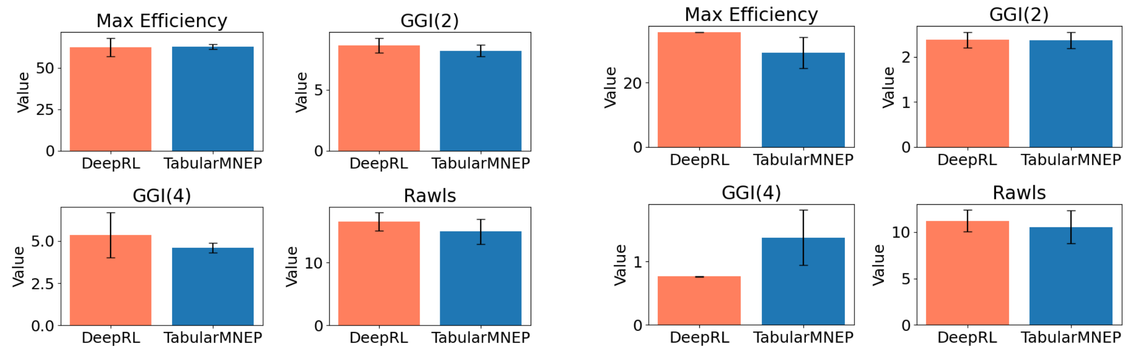
5 Experiments

We ran and evaluated the model in two real-world case study cities: Xi'an and Amsterdam. To facilitate introducing directional constraints and to provide higher granularity, both cities are split into grids of equally-sized cells, rather than relying on census tracts (this assumption can be relaxed).

Xi'an environment preparation

Wei et al. Wei et al. (2020) created and publicly released the Xi'an environment [2]. The city is organized into a $H^{29\times29}$ grid, comprising $1km^2$ cells. An origin-destination (OD) demand matrix was generated from GPS data collected over one month from 25 million mobile phones. Each cell is linked to an average house price index — we categorize them to five quintiles to create groups. We selected the average house price as a proxy for neighborhood development, as it is widely available across various cities and raises no privacy concerns. While our group definitions rely on this metric, they could also incorporate other attributes, such as those based on protected categories. The environment already includes two existing metro lines, and our experiments focus on expanding the network by designing a third line. This setting provides a wealth of mobility demand data, contrasting with the case study in Amsterdam discussed below.

[2] https://github.com/weiyu123112/City-Metro-Network-Expansion-with-RL



(a) Performance on the Xi'an environment.

(b) Performance on the Amsterdam environment.

Figure 3: We show results on the Xi'an environment. The proposed Tabular model performs on par with the Deep RL model across four diverse objectives in Xi'an (panel a) and Amsterdam (panel b).

Amsterdam environment preparation

The Amsterdam environment is organized into a $H^{35\times 47}$ grid of $0.5km^2$ cells. This cell size was chosen to maintain similar problem complexity in all cities, taking into account the smaller size of Amsterdam. Since GPS data are unavailable, we estimate the origin-destination (OD) demand using the recently published universal law of human mobility, which indicates that the total mobility flow between two areas i and j is determined by their distance and visitation frequency Schläpfer et al. (2021). We provide details on the estimation on Appendix B. As in the Xi'an environment, each cell is associated with an average house price sourced from the publicly available statistical bureau of the Netherlands [3]. The groups are defined as five quintiles based on this price.

5.1 Evaluation

We evaluate our proposed TabularMNEP algorithm against the state-of-the-art Deep Reinforcement Learning (DeepRL) method for Transport Network Design Wei et al. (2020), as well as a Genetic Algorithm (GA) Owais & Osman (2018) and an Ant-Colony Algorithm (ACA) Yang et al. (2007), with results for the latter two taken from Wei et al. (2020).

The methods are tested on four distinct reward functions: a utilitarian reward, maximizing total captured travel demand (Max Efficiency); two equal-sharing rewards using the Generalized Gini Index with weights of $1/2^i$ (GGI(2)) and $1/4^i$ (GGI(4)); and a Rawlsian reward that maximizes demand from the lowest house price quintile. We conducted a Bayesian hyperparameter search across 100 runs, selecting the top five configurations, running each five times, and choosing the one with the best average performance. DeepRL was trained over 3,500 epochs (128 episodes per epoch, totaling 448,000 episodes), while TabularRL required only 25,000 episodes—a reduction of 18-fold in total training episodes.

To estimate emissions (kg CO_2 equivalent), we consider GPU electricity consumption (kWh), total training hours, and the carbon emissions per kWh based on the 2024 monthly average for `COUNTRY`[4], using the formula: $\texttt{CO}_2 = \texttt{Watt} * \texttt{TrainingHours} * \texttt{CarbonFactor}$.

Model training used two types of in-house GPUs, the RTX 6000 Ada Generation (300 Watt) and GTX 1080Ti (250 Watt), depending on availability. Although our tabular method does not require a GPU, we report emissions based on GPU usage since a GPU-equipped node was reserved for model runs.

[3] https://www.cbs.nl/nl-nl/maatwerk/2019/31/kerncijfers-wijken-en-buurten-2019
[4] Country name retracted for anonymous submission.

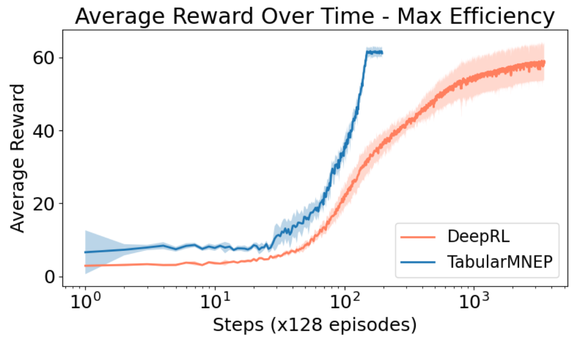


Figure 4: We demonstrate that the proposed TabularMNEP model achieves similar performance while requiring 18 times fewer episodes (x-axis is in log-scale).

	Xi'an				**Amsterdam**			
	Max. Efficiency	GGI(2)	GGI(4)	Rawls	Max. Efficiency	GGI(2)	GGI(4)	Rawls
GA (Wei et al. (2020))	28.9 ± 0.55	-	-	-	-	-	-	-
ACA (Wei et al. (2020))	30.3 ± 1.13	-	-	-	-	-	-	-
DeepRL	62.3 ± 5.66	8.7 ± 0.62	5.4 ± 1.33	16.6 ± 1.46	35.7 ± 0.06	2.4 ± 0.18	0.8 ± 0.00	11.2 ± 1.18
TabularMNEP (Ours)	62.6 ± 1.59	8.2 ± 0.46	4.6 ± 0.29	15.0 ± 1.98	29.4 ± 4.82	2.4 ± 0.18	1.4 ± 0.44	10.5 ± 1.80

Table 1: Results on Xi'an and Amsterdam for 5 seeds (GA and ACA are taken from Wei et al. (2020)).

6 Results

We ran both algorithms using 5 random seeds and provide code to replicate our results[5]. This section presents three key analyses: (1) a comparison of our proposed Tabular-TNDP method against recent approaches including Deep-RL, a Genetic Algorithm, and an Ant-Colony Algorithm (Section 6.1); (2) a demonstration of TabularMNEP's versatility across multiple social-good rewards (Section 6.2); and (3) a justification for choosing TabularMNEP in scenarios where interpretability is crucial (Section 6.3).

6.1 TabularMNEP performs on par with DeepRL methods

Figure 3 demonstrates that our proposed TabularMNEP method significantly outperforms both the Genetic Algorithm Owais & Osman (2018) and Ant-Colony Algorithm Yang et al. (2007) (GA and ACA results from Wei et al. (2020)). TabularMNEP achieves comparable performance to DeepRL across both the Xi'an and Amsterdam environments, considering both traditional and social good objectives defined in Section 3. Detailed averages and confidence intervals for all methods are presented in Table 1.

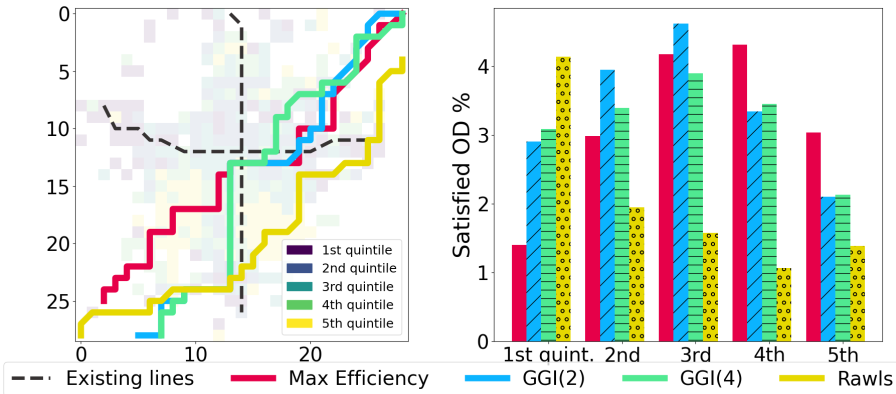
Notably, TabularMNEP achieves these results with substantially greater training efficiency, requiring only $25k$ episodes compared to DeepRL's $450k$ episodes (3500 epochs $\times$ 128 episodes). This $18\times$ reduction in training episodes is visualized in Figure 4 using a logarithmic x-axis.

In Table 2, we report the average CO_2 equivalent emissions from running our models across the four proposed reward functions. We observe that TabularMNEP requires, on average, $12\times$ fewer emissions to achieve performance comparable to the Deep RL baseline.

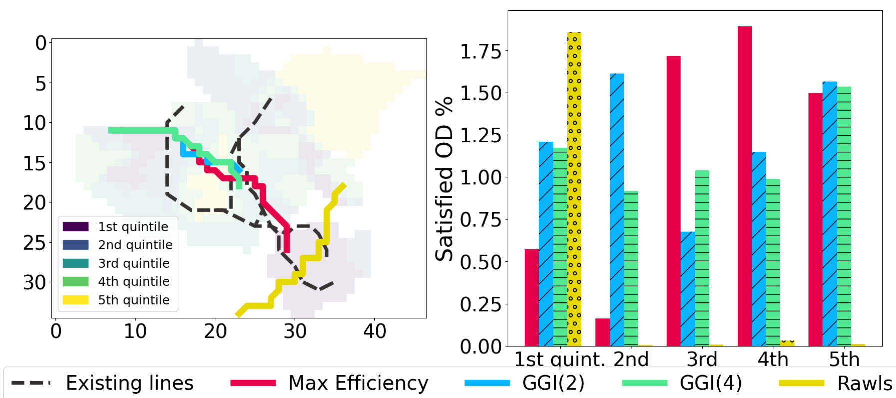
[5]https://github.com/*****/****

	Xi'an				**Amsterdam**			
	Max. Efficiency	GGI(2)	GGI(4)	Rawls	Max. Efficiency	GGI(2)	GGI(4)	Rawls
DeepRL	1.21	1.38	1.61	1.17	1.23	1.21	1.22	1.14
TabularMNEP (Ours)	0.05	0.13	0.13	0.12	0.06	0.28	0.28	0.10

Table 2: Estimated average emissions in kg CO_2 equivalent for each model's training for different reward functions.



(a) Generated Lines and Distribution of Benefits (Xi'an)



(b) Generated Lines and Distribution of Benefits (Amsterdam)

Figure 5: We present the results of applying various reward functions to design transportation lines in Xi'an (a) and Amsterdam (b). The left column displays the generated lines for each city, while the right column shows the distribution of satisfied demand across the five groups for the selected models.

6.2 TabularMNEP is capable of optimizing diverse rewards

Figure 5 shows the generated metro lines and the reward distribution among groups for both environments. The Max Efficiency reward function achieves the highest overall satisfied origin-destination flows, but we can observe that the rewards are distributed unequally among the five groups. In both Xi'an and Amsterdam, the highest quintiles exhibit greater satisfaction than the lowest quintiles, with inequality more pronounced in Amsterdam. This is due to the spatial distribution: in Xi'an, groups are more uniformly distributed, and segregation is lower, while in Amsterdam, the city center is dominated by higher-priced areas.

In contrast, the equality-based reward functions result in a more balanced distribution. Both GGI with $w = 2$ and $w = 4$ effectively equalize the rewards across groups. When $w = 4$, the rewards are distributed even more equally, though at the cost of overall efficiency. This allows decision-makers to control the trade-off between efficiency and fairness. The Rawls reward function prioritizes the lowest quintile in both environments, maximizing its satisfied demand. As intended, it directs the agent to optimize exclusively for the lowest quintile.

An additional insight from the Rawls reward function is its ability to reveal how isolated the lowest-utility group is. In Xi'an, maximizing for the lowest quintile creates "trickle-up" effects, benefiting other groups as well. However, in Amsterdam, where the lowest quintile is more segregated in the southeast, the generated line primarily benefits this group alone. This is further demonstrated in the spatial distribution of the lines, as shown in Figure 5.

6.3 TabularMNEP leads to more interpretable policies

Tabular RL offers a key advantage in solving the Metro Network Expansion Problem (MNEP): interpretability of the policies. As illustrated in Figure 6, we can visualize three critical aspects: (a) the optimal policy generating the metro line, (b) the average reward distribution across initial grid locations, and (c) the final Q-values with their corresponding best actions, which provide a direct interpretation for the best metro segment direction from each possible departing state. This interpretability provides decision-makers with insights beyond the model's output, enabling them to understand the relationship between actions and rewards, identify over-and under-explored areas in the city, and allowing genearting alternative routes to those produced by black-box models.

Transparency in this domain is particularly valuable as real-world metro planning often requires multiple alternative policies rather than a single solution. Additionally, tabular MNEP allows for incorporating spatial constraints after training, once the model has thoroughly explored the solution space. This post-training constraint application enables the model's ability to discover diverse solutions, while still capable of accommodating practical limitations.

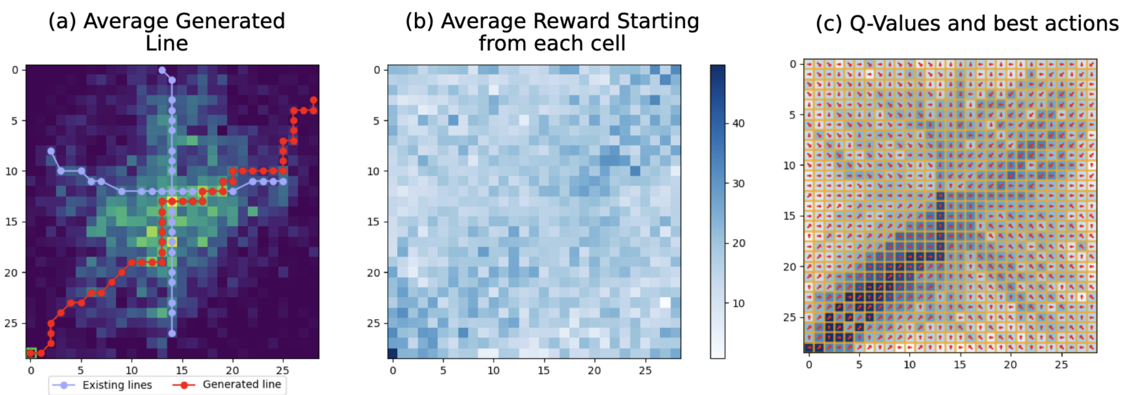


Figure 6: TabularRL not only effectively learns a diverse set of reward functions but also provides interpretability. In Panel (a), the metro line of a trained model optimized for maximum efficiency is illustrated. Panel (b) shows the average achievable reward from various starting points within the city, while Panel (c) displays the learned Q-values for each cell when the agent selects the action associated with the highest Q-value. Higher Q-values indicate more favorable locations for placing a metro station.

7 Conclusion

In this paper, we demonstrate that simple, tabular-based reinforcement learning methods can effectively address complex combinatorial optimization problems with diverse objectives, such as the Transport Network Design and Metro Network Expansion problems. Our approach involves reformulating the Markov Decision Process to create a smaller action space and creating distinct Q-tables for different action types.

We show that well-engineered problem reformulation combined with established methods can yield competitive results with significantly less computational power. Our method, which can run on standard personal computers without a GPU, performs comparably to state-of-the-art deep reinforcement learning methods, while using significantly less resources and training for significantly less time. Furthermore, this approach enhances interpretability and flexibility in policy selection. Consistent with recent research trends, our method showcases that effective computational policy-making in real-world applications is possible without relying on complex, black-box models. We hope this work will encourage a reevaluation of simpler models for other challenges too.

References

Zeke Ahern, Alexander Paz, and Paul Corry. Approximate multi-objective optimization for integrated bus route design and service frequency setting. *Transportation Research Part B: Methodological*, 155:1–25, January 2022. ISSN 0191-2615. doi: 10.1016/j.trb.2021.10.007. URL https://www.sciencedirect.com/science/article/pii/S0191261521001910.

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models, 2020. URL https://arxiv.org/abs/2007.03051.

Hamid Behbahani, Sobhan Nazari, Masood Jafari Kang, and Todd Litman. A conceptual framework to formulate transportation network design problem considering social equity criteria. *Transportation Research Part A: Policy and Practice*, 125:171–183, July 2019. ISSN 0965-8564. doi: 10.1016/j.tra.2018.04.005. URL https://www.sciencedirect.com/science/article/pii/S0965856417308030.

Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. Neural Combinatorial Optimization with Reinforcement Learning. *arXiv:1611.09940 [cs, stat]*, January 2017. URL http://arxiv.org/abs/1611.09940. arXiv: 1611.09940.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, April 2021. ISSN 03772217. doi: 10.1016/j.ejor.2020.07.063. URL https://linkinghub.elsevier.com/retrieve/pii/S0377221720306895.

Wenting Cheng, Jiahui Wu, William Moen, and Lingzi Hong. Assessing the spatial accessibility and spatial equity of public libraries' physical locations. *Library & Information Science Research*, 43(2):101089, April 2021. ISSN 0740-8188. doi: 10.1016/j.lisr.2021.101089. URL https://www.sciencedirect.com/science/article/pii/S0740818821000190.

Giuseppe Cuccu, Julian Togelius, and Philippe Cudre-Mauroux. Playing atari with six neurons, 2019. URL https://arxiv.org/abs/1806.01363.

Ahmed Darwish, Momen Khalil, and Karim Badawi. optimising Public Bus Transit Networks Using Deep Reinforcement Learning. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–7, Rhodes, Greece, September 2020. IEEE. ISBN 978-1-72814-149-7. doi: 10.1109/ITSC45102.2020.9294710. URL https://ieeexplore.ieee.org/document/9294710/.

Elizabeth Cahill Delmelle and Irene Casas. Evaluating the spatial equity of bus rapid transit-based accessibility patterns in a developing country: The case of Cali, Colombia. *Transport Policy*, 20:36–46, March 2012. ISSN 0967-070X. doi: 10.1016/j.tranpol.2011.12.001. URL https://www.sciencedirect.com/science/article/pii/S0967070X11001338.

Michel Deudon, Pierre Cournut, Alexandre Lacoste, Yossiri Adulyasak, and Louis-Martin Rousseau. Learning Heuristics for the TSP by Policy Gradient. In Willem-Jan van Hoeve (ed.), *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, Lecture Notes in Computer Science, pp. 170–181, Cham, 2018. Springer. ISBN 978-3-319-93031-2. doi: 10.1007/978-3-319-93031-2_12.

Ahmed El-Geneidy, David Levinson, Ehab Diab, Genevieve Boisjoly, David Verbich, and Charis Loong. The cost of equity: Assessing transit accessibility and social disparity using total travel cost. *Transportation Research Part A: Policy and Practice*, 91:302–316, September 2016. ISSN 0965-8564. doi: 10.1016/j.tra.2016.07.003. URL https://www.sciencedirect.com/science/article/pii/S0965856416305924.

Wei Fan and Randy B. Machemehl. Using a Simulated Annealing Algorithm to Solve the Transit Route Network Design Problem. *Journal of Transportation Engineering*, 132(2):122–132, February 2006. ISSN 0733-947X. doi: 10.1061/(ASCE)0733-947X(2006)132:2(122). URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%290733-947X%282006%29132%3A2%28122%29. Publisher: American Society of Civil Engineers.

Reza Zanjirani Farahani, Elnaz Miandoabchi, W. Y. Szeto, and Hannaneh Rashidi. A review of urban transportation network design problems. *European Journal of Operational Research*, 229(2):281–302, September 2013. ISSN 0377-2217. doi: 10.1016/j.ejor.2013.01.001. URL https://www.sciencedirect.com/science/article/pii/S0377221713000106.

Steven Farber, Keith Bartholomew, Xiao Li, Antonio Páez, and Khandker M. Nurul Habib. Assessing social equity in distance based transit fares using a model of travel behavior. *Transportation Research Part A: Policy and Practice*, 67:291–303, September 2014. ISSN 0965-8564. doi: 10.1016/j.tra.2014.07.013. URL https://www.sciencedirect.com/science/article/pii/S0965856414001785.

Valérie Guihaire and Jin-Kao Hao. Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice*, 42(10):1251–1273, 2008.

Gabriel Gutiérrez-Jarpa, Gilbert Laporte, and Vladimir Marianov. Corridor-based metro network design with travel flow capture. *Computers & Operations Research*, 89:58–67, January 2018. ISSN 0305-0548. doi: 10.1016/j.cor.2017.08.007. URL https://www.sciencedirect.com/science/article/pii/S0305054817302137.

Diego Hernandez. Uneven mobilities, uneven opportunities: Social distribution of public transport accessibility to jobs and education in Montevideo. *Journal of Transport Geography*, 67:119–125, 2018. ISSN 0966-6923. doi: 10.1016/j.jtrangeo.2017.08.017. URL https://www.sciencedirect.com/science/article/pii/S0966692316303556.

Sami Jullien, Mozhdeh Ariannezhad, Paul Groth, and Maarten de Rijke. A simulation environment and reinforcement learning method for waste reduction. *arXiv preprint arXiv:2205.15455*, 2022.

Wouter Kool, Herke van Hoof, and Max Welling. Attention, learn to solve routing problems! In *International Conference on Learning Representations*, 2018.

Srivatsan Krishnan, Maximilian Lam, Sharad Chitlangia, Zishen Wan, Gabriel Barth-Maron, Aleksandra Faust, and Vijay Janapa Reddi. Quarl: Quantization for fast and environmentally sustainable reinforcement learning, 2022. URL https://arxiv.org/abs/1910.01055.

Gilbert Laporte and Marta M. B. Pascoal. Path based algorithms for metro network design. *Computers & Operations Research*, 62:78–94, October 2015. ISSN 0305-0548. doi: 10.1016/j.cor.2015.04.007. URL https://www.sciencedirect.com/science/article/pii/S0305054815000878.

Marcy Lowe, Bengu Aytekin, and Gary Gereffi. Public Transit Buses: A Green Choice Gets Greener. October 2009.

Karel Martens. *Transport Justice: Designing fair transportation systems*. Routledge, July 2016. ISBN 978-1-317-59958-6. Google-Books-ID: m0yTDAAAQBAJ.

Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 134:105400, October 2021. ISSN 0305-0548. doi: 10.1016/j.cor.2021.105400. URL https://www.sciencedirect.com/science/article/pii/S0305054821001660.

Dimitris Michailidis, Willem Röpke, Sennay Ghebreab, Diederik M Roijers, and Fernando P Santos. Fairness in Transport Network Design - A Multi-Objective Reinforcement Learning Approach. *Adaptive and Learning Agents Workshop*, 2023.

Herbert Mohring. Optimization and scale economies in urban bus transportation. *The American Economic Review*, 62(4):591–604, 1972.

Muhammad Ali Nayeem, Md Monirul Islam, and Xin Yao. Solving transit network design problem using many-objective evolutionary approach. *IEEE Transactions on Intelligent Transportation Systems*, 20(10): 3952–3963, 2018.

MohammadReza Nazari, Afshin Oroojlooy, Lawrence Snyder, and Martin Takac. Reinforcement learning for solving the vehicle routing problem. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/9fb4651c05b2ed70fba5afe0b039a550-Paper.pdf`.

Grigory Neustroev, Sytze P. E. Andringa, Remco A. Verzijlbergh, and Mathijs M. De Weerdt. Deep Reinforcement Learning for Active Wake Control. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 944–953, Richland, SC, May 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-9213-6.

Mahmoud Owais and Mostafa K. Osman. Complete hierarchical multi-objective genetic algorithm for transit network design problem. *Expert Systems with Applications*, 114:143–154, December 2018. ISSN 0957-4174. doi: 10.1016/j.eswa.2018.07.033. URL `https://www.sciencedirect.com/science/article/pii/S0957417418304573`.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training, 2021. URL `https://arxiv.org/abs/2104.10350`.

Rafael H. M. Pereira, David Banister, Tim Schwanen, and Nate Wessel. Distributional effects of transport policies on inequalities in access to opportunities in Rio de Janeiro. *Journal of Transport and Land Use*, 12 (1):741–764, 2019. ISSN 1938-7849. URL `https://www.jstor.org/stable/26911287`. Publisher: Journal of Transport and Land Use.

V. Dimitra Pyrialakou, Konstantina Gkritza, and Jon D. Fricker. Accessibility, mobility, and realized travel behavior: Assessing transport disadvantage from a policy perspective. *Journal of Transport Geography*, 51:252–269, February 2016. ISSN 0966-6923. doi: 10.1016/j.jtrangeo.2016.02.001. URL `https://www.sciencedirect.com/science/article/pii/S0966692316000144`.

Govardana Sachithanandam Ramachandran, Ivan Brugere, Lav R. Varshney, and Caiming Xiong. GAEA: Graph Augmentation for Equitable Access via Reinforcement Learning. *arXiv:2012.03900 [cs]*, April 2021. URL `http://arxiv.org/abs/2012.03900`. arXiv: 2012.03900.

Naveen Raman, Sanket Shah, and John Dickerson. Data-Driven Methods for Balancing Fairness and Efficiency in Ride-Pooling. *arXiv:2110.03524 [cs]*, October 2021. URL `http://arxiv.org/abs/2110.03524`. arXiv: 2110.03524.

Markus Schläpfer, Lei Dong, Kevin O'Keeffe, Paolo Santi, Michael Szell, Hadrien Salat, Samuel Anklesaria, Mohammad Vazifeh, Carlo Ratti, and Geoffrey B. West. The universal visitation law of human mobility. *Nature*, 593(7860):522–527, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03480-9. URL `https://www.nature.com/articles/s41586-021-03480-9`. Number: 7860 Publisher: Nature Publishing Group.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning Fair Policies in Multiobjective (Deep) Reinforcement Learning with Average and Discounted Rewards. *arXiv:2008.07773 [cs]*, August 2020. URL `http://arxiv.org/abs/2008.07773`. arXiv: 2008.07773.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, Apr. 2020. doi: 10.1609/aaai.v34i09.7123. URL `https://ojs.aaai.org/index.php/AAAI/article/view/7123`.

Hongyuan Su, Yu Zheng, Jingtao Ding, Depeng Jin, and Yong Li. Metrognn: Metro network expansion with reinforcement learning. In *Companion Proceedings of the ACM on Web Conference 2024*, pp. 650–653, 2024.

W. Y. Szeto and Y. Jiang. Transit route and frequency design: Bi-level modeling and hybrid artificial bee colony algorithm approach. *Transportation Research Part B: Methodological*, 67:235–263, 2014. ISSN

0191-2615. doi: 10.1016/j.trb.2014.05.008. URL https://www.sciencedirect.com/science/article/pii/S0191261514000812.

David Tedjopurnomo, Zhifeng Bao, Farhana Choudhury, Hui Luo, and A. K. Qin. Equitable Public Bus Network Optimization for Social Good: A Case Study of Singapore. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 278–288, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533092. URL https://dl.acm.org/doi/10.1145/3531146.3533092.

Anne S. van der Veen, Jan Anne Annema, Karel Martens, Bart van Arem, and Gonçalo Homem de Almeida Correia. Operationalizing an indicator of sufficient accessibility – a case study for the city of Rotterdam. *Case Studies on Transport Policy*, 8(4):1360–1370, 2020. ISSN 2213-624X. doi: 10.1016/j.cstp.2020.09.007. URL http://www.sciencedirect.com/science/article/pii/S2213624X20301024.

Bert van Wee. Discussing Equity and Social Exclusion in Accessibility Evaluations. pp. 18, 2011.

Lebing Wang, Jian Gang Jin, Gleb Sibul, and Yi Wei. Designing Metro Network Expansion: Deterministic and Robust Optimization Models. *Networks and Spatial Economics*, 23(1):317–347, March 2023. ISSN 1572-9427. doi: 10.1007/s11067-022-09584-7. URL https://doi.org/10.1007/s11067-022-09584-7.

Qi Wang and Chunlei Tang. Deep reinforcement learning for transportation network combinatorial optimization: A survey. *Knowledge-Based Systems*, 233:107526, December 2021. ISSN 0950-7051. doi: 10.1016/j.knosys.2021.107526. URL https://www.sciencedirect.com/science/article/pii/S0950705121007887.

Yu Wei, Minjia Mao, Xi Zhao, Jianhua Zou, and Ping An. City Metro Network Expansion with Reinforcement Learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2646–2656, Virtual Event CA USA, August 2020. ACM. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403315. URL https://dl.acm.org/doi/10.1145/3394486.3403315.

Ziyi Xu, Xue Cheng, and Yangbo He. Performance of Deep Reinforcement Learning for High Frequency Market Making on Actual Tick Data. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 1765–1767, Richland, SC, May 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-9213-6.

Zhongzhen Yang, Bin Yu, and Chuntian Cheng. A parallel ant colony algorithm for bus network optimization. *Computer-Aided Civil and Infrastructure Engineering*, 22(1):44–55, 2007.

Amirali Zarrinmehr, Mahmoud Saffarzadeh, Seyedehsan Seyedabrishami, and Yu Marco Nie. A path-based greedy algorithm for multi-objective transit routes design with elastic demand. *Public Transport*, 8:261–293, 2016.

Liqing Zhang, Leong Hou U, Shaoquan Ni, Dingjun Chen, Zhenning Li, Wenxian Wang, and Weizhi Xian. City metro network expansion based on multi-objective reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 169:104880, 2024. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2024.104880. URL https://www.sciencedirect.com/science/article/pii/S0968090X24004017.

A Appendix – Feasibility Rules

The feasibility rules applied in this paper closely resemble those in previous studies Wei et al. (2020); Zhang et al. (2024). The agent's actions are constrained using an `ActionMask`, which is updated at each timestep based on the agent's current location and prior positions. This approach ensures that the agent moves forward, avoids cyclical paths, and does not revisit locations where a station has already been placed.

Our method optimizes this process by maintaining a constant action mask length of 8, representing all possible directions (including diagonals), rather than the entire grid size. The agent's movement direction is established by its initial longitudinal and latitudinal steps. For example, if the agent begins by moving north, southward actions will be masked out to enforce forward progression. If the agent subsequently moves east, only actions corresponding to the north, east, and northeast directions remain available, with all other actions masked. Figure 7 illustrates how these feasibility rules are applied through the action mask during an episode.

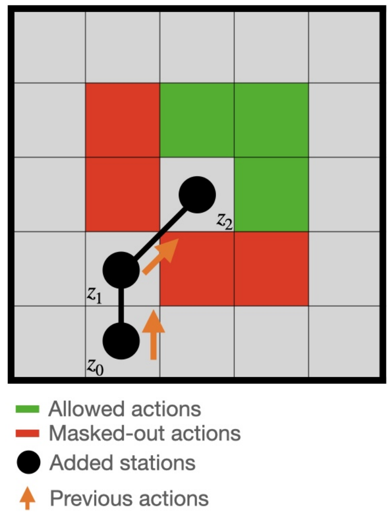


Figure 7: A snapshot of an episode, where the action mask created by feasibility rules constraints the next available actions to the agent.

B Appendix – Amsterdam environment preparation

GPS data is unavailable for Amsterdam, so we estimate the origin-destination (OD) demand using the recently published universal law of human mobility, which indicates that the total mobility flow between two areas i and j is determined by their distance and visitation frequency Schläpfer et al. (2021). The calculation is as follows:

$$OD_{ij} = \mu_j \mathsf{K_i} / d_{ij}^2 \ln(f_{max}/f_{min}) \tag{7}$$

Here, $\mathsf{K_i}$ is the total area of the origin location i, d_{ij}^2 is the (Manhattan) distance between i and j, and μ_j represents the magnitude of flows, computed as:

$$\mu_j \approx \rho_{pop}(j) rad_j^2 f_{max} \tag{8}$$

Where rad_j^2 is the radius of area j. We estimate the flows over a week by setting f_{min}, f_{max} to 1/7 and 7 respectively. The grid cells are of equal size, K_i and can be omitted from the calculation.