City Metro Network Expansion with Reinforcement Learning

Yu Wei
School of Electronic and Information Engineering, Xi'an Jiaotong University
weiyu123112@163.com

Minjia Mao
School of Mathematics and Statistics, Xi'an Jiaotong University
maominjia@foxmail.com

Xi Zhao*†
School of Management, Xi'an Jiaotong University
Zhaoxi1@mail.xjtu.edu.cn

Jianhua Zou‡
School of Electronic and Information Engineering, Xi'an Jiaotong University
jhzou@sei.xjtu.edu.cn

Ping An
School of Management, Xi'an Jiaotong University
sdzx119@163.com

ABSTRACT

City metro network expansion, included in the transportation network design, aims to design new lines based on the existing metro network. Existing methods in the field of transportation network design either (i) can hardly formulate this problem efficiently, (ii) depend on expert guidance to produce solutions, or (iii) appeal to problem-specific heuristics which are difficult to design. To address these limitations, we propose a reinforcement learning based method for the city metro network expansion problem. In this method, we formulate the metro line expansion as a Markov decision process (MDP), which characterizes the problem as a process of sequential station selection. Then, we train an actor-critic model to design the next metro line on the basis of the existing metro network. The actor is an encoder-decoder network with an attention mechanism to generate the parameterized policy which is used to select the stations. The critic estimates the expected cumulative reward to assist the training of the actor by reducing training variance. The proposed method does not require expert guidance during design, since the learning procedure only relies on the reward calculation to tune the policy for better station selection. Also, it avoids the difficulty of heuristics designing by the policy formalizing the station selection. Considering origin-destination (OD) trips and social equity, we expand the current metro network in Xi'an, China, based on the real mobility information of 24,770,715 mobile phone users in the whole city. The results demonstrate the advantages of our method compared with existing approaches.

*Xi Zhao is the first corresponding author of this paper.
†Also with the Key Lab of the Ministry of Education for Process Control & Efficiency Engineering, Xi'an, 710049, China.
‡Jianhua Zou is the second corresponding author of this paper.

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403315

CCS CONCEPTS

• **Computing methodologies** → **Planning and scheduling**; Reinforcement Learning; • **Transportation network design** → Metro network expansion.

KEYWORDS

Metro network expansion; Reinforcement learning; Actor-critic model; Social equity

ACM Reference Format:
Yu Wei, Minjia Mao, Xi Zhao, Jianhua Zou, and Ping An. 2020. City Metro Network Expansion with Reinforcement Learning. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3394486.3403315

1 INTRODUCTION

The city metro network plays an important role in the public transportation system. As the city has developed, new transportation demands have led to the expansion of the metro network. The last few years have witnessed a tremendous expansion of the metro network [27]. Meanwhile, the expansion of the metro network in turn has a profound impact on the city. Expanded lines may change the mobility trend of city population. Metro network expansion and city dynamics are mutually influenced, therefore it is more reasonable to expand the metro network by gradually designing new lines according to the latest city dynamic. In this study, we design the next metro line, consisting of stations and line routing, to expand the existing metro network. This process can be conducted repeatedly to achieve a multi-line metro network expansion [4, 20].

Metro network expansion is included in the transportation network design. Usually, the objective of transportation network design is mobility-based, such as maximizing satisfied OD trips [21]. As society progresses, sustainability has increasingly become the demand of city development. The need for sustainability prompts governments to realize other impacts of the transportation system, and thereby influences their transport policy [24]. Among these impacts, the importance of social equity [2] has been acknowledged, and there have been several realistic transportation plans considering social equity [1]. The metro network, an important city transportation system, has a great influence on social equity.

Therefore, in this work, we consider both OD trips and social equity to expand the city metro network.

There are various studies dealing with transportation network design problems [8, 15], and they can be mainly divided into two categories. One category is based on mathematical programming. Several studies [11, 12, 22, 30] formulate transportation network design problems as non-linear integer programming models, and adopt solvers to obtain solutions. However, their formulations call for an exponential number of shape constraints to ensure the rationality of the transportation network, which hinders solving the problem efficiently [5]. Besides the formulations, their solution technologies are intractable for a realistic sized problem. To make large-scale problems solvable, they predefine corridors based on expert knowledge to limit the search space, and only consider designing the transportation network in these corridors. Their results depend on expert guidance, and the best solution may be left out. In summary, these studies can hardly formulate the transportation network design problem appropriately [5, 19], and their solutions heavily depend on expert guidance, which lacks reliability.

Another category alternative to mathematical programming is based on heuristics. Owais *et al.* [26] utilize Genetic Algorithm to generate a bus route network from an existing transportation network. Dufourd *et al.* [6] design a tabu search heuristic for locating a rapid transit line. Yang *et al.* [32] propose an ant colony algorithm to design a bus network, with a specification of origin and destination by experts. For different transportation network design problems, these studies design problem-specific heuristics to obtain the solutions. However, the problem-specific heuristics can be difficult to design, especially in cases like a metro network with rigorous shape constraints. The commonly used operators in Genetic Algorithm [26] and Tabu Search [6], which focus on the connectivity of the route, are likely to lead to an infeasible metro line, which makes these methods inefficient for designing a metro network, and provides no guarantee as to the quality of solutions. Therefore, designing methods without heuristics is an urgent need.

Considering all of the above, existing methods in transportation network design can hardly be applied to the metro network expansion problem. Instead, we need an efficient formulation and a generic method to solve the metro network expansion problem without expert guidance.

In this paper, we propose a RL based method to solve the city metro network expansion problem. We consider the metro line expansion as a process of sequential station selection, and then naturally formulate this process as a Markov decision process (MDP). To ensure reasonable connection patterns between stations of a metro line, we also design feasibility rules based on the selected station sequence for the next station selection. This formulation efficiently characterizes the expansion of the metro line, without heavy constraints like existing studies [30].

Following this formulation, we propose an actor-critic model [17] to generate the next metro line. The actor adopts an encoder-decoder network with an attention mechanism to represent the parameterized policy, which maps the current metro network state to a probability distribution for station selection. Specifically, in the actor, an encoder characterizes the timely metro station information in the expansion process, an RNN decoder characterizes the sequence information of the selected stations, and an attention layer [29] integrates these two sets of information to produce a probability distribution over feasible candidate stations. In the critic, a neural network is used to estimate the expected cumulative reward. Only requiring the reward calculation, we employ a policy gradient algorithm [10] to train our network to find a high-priority metro line. Without expert guidance, the learning procedure drives the policy to keep track of superior solutions during the search and to find better solutions. Its natural exploration mechanism determines that our method is suitable for large scale solution space. The parameterized policy formalizes the station selection, avoiding the difficulties of heuristics design.

Using the real city-scale human mobility information of 24,770,715 mobile phone users obtained from a citywide 3G cellular network, we expand the existing metro network in Xi'an, China. The results demonstrate the effectiveness of our method. Our contributions are as follows:

- We incorporate social equity concerns with mobility demands into metro network expansion. By proposing a weighted sum reward construction, our RL method can take multi-factors into consideration.
- We formulate the metro line expansion problem as a Markov decision process, and design feasibility rules based on the selected station sequence to ensure the reasonable connection patterns of the metro line, which is a more efficient formulation method than integer programming models.
- We are the first to propose a RL based method to solve the city metro network expansion problem. With the exploration mechanism of RL, our method can generate solutions without expert guidance.
- We use real city-scale human mobility information to expand a metro network. The experimental results demonstrate the effectiveness of our method.

2 RELATED WORK

2.1 Transportation Network Design

Several studies [8, 15] have reviewed the transportation network design literature, and these existing methods mainly fall into two categories, mathematical programming methods and heuristics methods. Mathematical programming methods formulate this problem as nonlinear integer programming models, and usually obtain the solutions using a solver. Gutiérrez-Jarpa *et al.* [11] first select a set of corridors with higher passenger traffic by using greedy generation heuristics, and then consider designing metro lines in these predefined corridors. Wei *et al.* [30] predefine corridors and introduce a bi-objective model to expand the metro network in Wuxi, China. However, their huge constraints, which ensure the rationality of the transportation network, lead to an ineffective solution method in a large-scale space, unless based on expert guidance to predefine the corridors.

As for the second category, search-based heuristic methods, such as Simulated Annealing [7] and Tabu Search [6], first generate the initial solutions, and then modify the initial solutions with the help of heuristics to get better solutions. Genetic Algorithm [26] randomly generates initial routes, and then design operators to evolve routes for better solutions. However, the commonly used heuristic

operators in transportation networks may lead to infeasible solutions for the metro expansion problem. It is hard to come out with appropriate heuristics for problems with heavy constraints like metro expansion.

2.2 Reinforcement Learning

The strength of RL lies in its powerful decision-making ability. RL has made great progress in complicated tasks like playing Atari games [25], recommender systems [35], combinatorial optimization [3] and so on. Existing RL methods can be divided into three categories [10]: actor-only, critic-only, and actor-critic methods. With regard to metro network expansion, we find that it can be formulated as a sequential decision-making process. Then, through technology combing, we finally employ an actor-critic method to expand the metro network.

3 PROBLEM DEFINITION

In this paper, we design the next metro line to expand the current metro network in a target city. The metro line is determined by stations and arcs connecting the stations, and is allowed to intersect with existing lines to form transfer stations. We define the metro expansion problem as follows.

For a target city, we divide it into $n \times n$ grids in a two-dimensional space $\{g_i\}_{i=0}^{n^2-1}$. Each grid g_i is a square with a width of d_0, and its center is a candidate station i. We define the expansion of the metro network on an undirected graph $G = (\mathbb{N}, \mathbb{E})$, where $\mathbb{N} = \{0, 1, ..., n^2 - 1\}$ contains all candidate stations and $\mathbb{E} = \{(i, j) : i, j \in \mathbb{N}\}$ contains all edges which directly connect the stations i and j. Among the G, several nodes and the edges connecting these nodes form the existing metro network, and these nodes are the candidate transfer stations connecting existing lines and the newly-built line. Each grid g_i is associated with a compound index of development D_i, and any two candidate stations i and j are associated with a travel flow capture, which contains the total OD trips starting at one of the two stations i and j and ending at another. We denote the travel flow capture between station i and j as $od_{i,j} (= od_{j,i})$.

We present the expanded metro line as an ordered station sequence $Z = (z_1, z_2, \ldots, z_T), z_i \in \mathbb{N}$, where the adjacent stations are directly connected. In practice, the expanded line Z should satisfy the following constraints:

- The consecutive stations must follow the minimum-maximum distance rules [22]. That is to say, the separation between two consecutive stations $d(z_i, z_{i+1})$ must satisfy $d_{min} \leq d(z_i, z_{i+1}) \leq d_{max}, i \in \{1, 2, \ldots, T-1\}$, where d_{min} and d_{max} are the minimum and maximum separation between any two consecutive stations.
- The line shape should ensure reasonable connection patterns between stations, avoiding sub-tour and squiggly lines [30].
- The number of the stations T is limited by N.
- The budget for construction is limited by B_0.

Based on the above, the newly satisfied OD trips by the expanded line Z are defined as follows:

$$R_{od}(Z) = \sum_i \sum_j od_{i,j} + \sum_i \sum_k x_{i,k} \times od_{i,k}, \tag{1}$$

where $i, j (i < j)$ are the stations on the expanded line Z, $k (k \neq i)$ is the station on existing lines and not on Z, and $x_{i,k}$ is set to 1 if there is a path connecting the two stations i and k along the metro network; otherwise it is set to 0. In Equation (1), the first term presents the direct OD trips achieved by Z, and the second term presents the OD trips between Z and existing lines through transfer stations.

As to the social equity, we consider the distributable benefits for the grids traversed by the expanded line Z. According to Appendix C, the social equity indicator of Z is calculated as:

$$R_{ac}(Z) = \sum_i Ac_{g_i}, \tag{2}$$

where i are the stations on the expanded line Z, and Ac_{g_i} referring to Equation (15) is the accessibility index of grid g_i.

In our study, the objective is to design the next metro line Z based on the current metro network to maximize the satisfied transportation demands which are defined as the weighted sum of newly satisfied OD trips and social equity

$$\omega(Z|G) = \alpha_1 \times R_{od}(Z) + \alpha_2 \times R_{ac}(Z), \tag{3}$$

where G is the underlying network that defines the problem, α_1 and α_2 are the weights of added OD trips $R_{od}(Z)$ and social equity $R_{ac}(Z)$, and $\alpha_1 + \alpha_2 = 1$.

4 METHOD

According to the definition in Section 3, along a metro line, a preceding station determines the area to locate its succeeding station, and the preceding section determines the layout of the succeeding section. The subsequent station locating is influenced by the previous stations, so that the generation of a metro line can be viewed as a process of sequential decisions about where to locate the stations. In our study, we consider metro line expansion as a sequential station selection process, and leverage RL to optimize the sequential decisions to obtain the next metro line.

4.1 RL Formulation

In this section, we formulate the metro line expansion as a MDP. Taking an ordered station sequence $Z = (z_1, z_2, \ldots, z_T), z_i \in \mathbb{N}$ generated during an episode as an example, the elements of this MDP are as follows:

- State space $\mathcal{S}$. A state $s_t \in \mathcal{S}$ is defined to characterize the selected station sequence $Z_{t-1} = (z_1, z_2, \ldots, z_{t-1})$ before step t, where $t = \{1, .., T, T+1\}$ and Z_0 is an empty set.
- Action space $\mathcal{A}$. The action $a_t \in \mathcal{A}$ is defined as the station $z_t \in \mathbb{N}$ selected at step t.
- A deterministic state transition function $p(s_{t+1}|s_t, a_t)$. When the agent selects an action a_t at state s_t, the transition function determines the next state $s_{t+1} : Z_t = (z_1, z_2, \ldots, z_t)$.
- Reward function $r(s_t, a_t)$. We expect the expanded metro line Z to achieve more transportation demands, thus we restrict our attention to the objective value $\omega(Z|G)$ of a complete sequence (refer to Equation (3) for more details). In our study, the reward function is set to 0 if $s_t (t < T+1)$ is not a terminal state; otherwise it is set to $\omega(Z|G)$ at the terminal step $T+1$.

Within the MDP, a parameterized policy $\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, which maps states to a probability distribution over the actions, is used

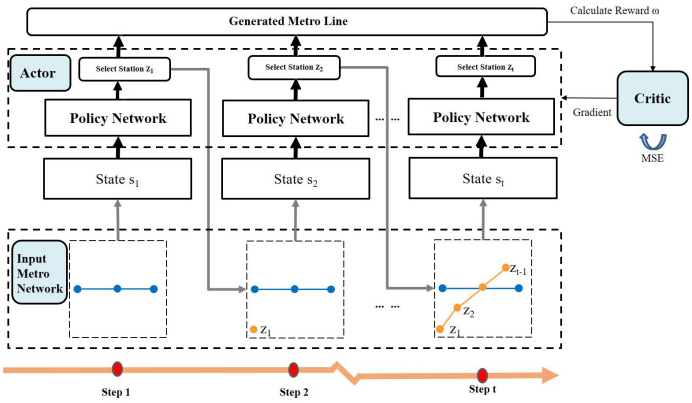


Figure 1: Framework. The bule (orange) line in the input is the existing (expanded) metro line.

to select actions. Here $\mathcal{P}(\mathcal{A})$ is the set of probability of selecting each action and θ is the adjustable parameters.

In addition to the policy π_θ, in order to ensure the metro line achieves the constraints in Section 3, we also design feasibility rule $F(s_t, G)$, which is based on the agent selecting station sequence and the existing metro network, to determine the optional actions in step t (see Appendix A for more details about agent interaction with the environment). Thus, based on the policy $\pi_\theta(a_t|s_t)$ and feasibility rules $F(s_t, G)$, the agent interacts with the environment to select the action as follows:

$$p(a_t|s_t, G) = \pi_\theta(a_t|s_t) \circ F(s_t, G), \tag{4}$$

where $\circ$ represents the generalized operator of two functions (see Section 4.3 for more details).

Now, for a metro line Z, its generating probability, according to the probability chain rule, is

$$p(Z|G) = \prod_{t=1}^{T} p(z_t|Z_{t-1}, G), \tag{5}$$

where $p(z_t|Z_{t-1}, G)$ (calculated as Equation (4)) presents the probability of selecting the next station z_t based on selected station sequence Z_{t-1} at step t. This metro line Z is associated with a cumulative return $\omega(Z|G)$, based on the definition of our reward function $r(s_t, a_t)$.

Our goal is to find a policy π_θ to maximize the expected cumulative reward which, given a network G, is defined as

$$J(\theta|G) = \mathbb{E}_{Z\sim\pi_\theta}\omega(Z|G). \tag{6}$$

4.2 Framework

According to the RL formulation in Section 4.1, the dimension of the action space in our metro network expansion problem is related to the number of stations. For a realistic size problem, the high-dimensional action space causes value-based methods, such as DQN, to be less suitable [23]. Thus, we employ policy-based technology to solve the problem, which directly parameterizes the policy. Figure 1 depicts our actor-critic framework. The actor takes the metro network as input, and outputs the expanded metro line. Its core component is a policy network, which maps metro network states to the action space for the station selection, one station at each step. By concatenating these selected stations in order, the expanded metro line is generated. The reward is calculated with the generated metro line. The critic estimates the expected cumulative reward to assist the training of the actor by reducing training variance. Next, we elaborate the policy network in the actor.

4.3 Policy Network Architecture

During the generation of a metro line, the selected stations affect the subsequent optional stations, which is guaranteed by the feasibility rules in Section A. We expect the policy to capture the dependencies between stations that can coexist to generate a feasible metro line, and give high probability to the dependent station sequences that achieve more transportation demands.

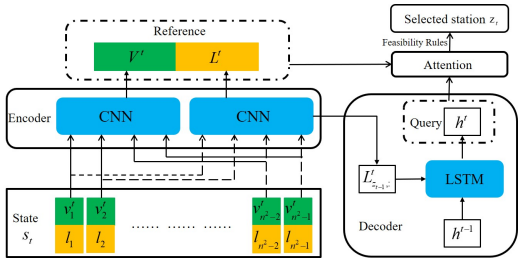


Figure 2: Policy network. It takes the state s_t as input, and generate the probability distribution to select a station.

To achieve the above requirements, we employ an encoder-decoder neural network coupling with an attention layer to parameterize our station selection policy, as shown in Figure 2. The policy network takes the metro network state as input, and outputs the selected station, one station at a step. Specifically, the encoder creates the representations for the metro network, one representation at a step. The decoder employs an RNN to characterize the sequence information of the selected stations during the generation of a metro line. Taking these two sets of information, an attention layer [29], which is able to flexibly model the dependencies between stations without regard to their distance in a station sequence [28], generates a probability distribution over all stations to guide the station selection, one station at a step. Next, we elaborate on how our policy network works.

At each step t, the state s_t characterizes the timely metro network information, aiming to distinguish the state changes caused by agent actions. To achieve this, we represent each station i by a tuple $X_i^t = \left(l_i, v_i^t\right)$, where l_i is the two-dimensional coordinates of station

i in the grid space, and the natural number $v_i^t \in [0, t-1]$ indicates this station i is selected at step v_i^t ($v_i^t = 0$ means that this station has not been selected before step t). Then, the state s_t is the sequential concatenation of the tuples of all stations $s_t = \{X_0^t, X_1^t, \ldots, X_{n^2-1}^t\}$, which consists of the location feature and the station selection feature. Here the station number starts at 0 and ends at $n^2 - 1$.

With this input state s_t, the encoder computes the embeddings for the stations through two 1-dimensional convolutional neural networks respectively, each with d filters, one for the location feature and another for the station selection feature. We denote these two embedded features as $\mathbf{L}^t$ and $\mathbf{V}^t$, and these two embedded features have a common dimension d for each station.

Next, the decoding happens. The decoder consists of a Long Short Term Memory (LSTM) [14]. Again, we use the ordered station sequence $Z = (z_1, z_2, \ldots, z_T), z_i \in \mathbb{N}$ in Section 4.1 as an example. At step t, it takes its own hidden state $\mathbf{h}^{t-1}$ at step $t-1$ and the embedded location feature $\mathbf{L}^t_{z_{t-1},:}$ of the last selected station z_{t-1} as inputs to generate the current hidden state $\mathbf{h}^t$ as follows:

$$\mathbf{h}^t, \mathbf{c}^t = LSTM\left(\mathbf{L}^t_{z_{t-1},:}, \mathbf{h}^{t-1}, \mathbf{c}^{t-1}\right), \tag{7}$$

where $\mathbf{c}^t$ presents the cell state of LSTM itself at step t. When $t = 1$, $\mathbf{h}^0$ and $\mathbf{c}^0$ are initialized to the zero tensor with dimension d.

After decoding, the attention layer takes the embedded features of stations as reference and the hidden state $\mathbf{h}^t$ as query to generate the probability distribution over total stations. Specifically, at step t, the attention mechanism is modeled as follows:

$$q_i^t = \mathbf{v}_a^T \tanh\left(\mathbf{W}_a\left(\mathbf{L}^t_{i,:} + \mathbf{V}^t_{i,:} + \mathbf{h}^t\right)\right), i \in \left\{0, 1, \cdots, n^2-1\right\} \tag{8}$$

$$\mathbf{u}^t = \mathbf{q}^t \oplus (H \cdot F(s_t, G)), \tag{9}$$

$$p\left(z_t | Z_{t-1}, G\right) = softmax\left(\mathbf{u}^t\right), \tag{10}$$

where softmax normalizes the vector $\mathbf{u}^t$ to generate a probability distribution over total stations, $\mathbf{L}^t_{i,:}$ and $\mathbf{V}^t_{i,:}$ are the embedded features of station i, $\mathbf{h}^t$ is the hidden state of the decoder, the vector $\mathbf{q}^t$ is the sequential concatenation of attention score q_i^t, H is a huge constant, the binary vector $F(s_t, G)$ reflects the feasibility rules, $\oplus$ represents the element-wise sum operator, and the matrix $\mathbf{W}_a$ and the vector $\mathbf{v}_a$ are training parameters.

At step t, the agent selects the station z_t according to the probability distribution in Equation (10). Then, the metro network state and the feasibility rules change accordingly. The policy network takes these as inputs for the next station selection. The process of choosing stations is repeated until any termination condition is reached (see Appendix A for more details about the agent interaction with the environment). The generation process of a metro line is presented in Algorithm 1.

4.4 Actor-Critic Training

In this section, we aim to train the policy network which is parameterized with parameters θ to maximize the expected cumulative reward in Equation (6). We use the policy gradient based actor-critic algorithm [10], in which the actor is synonymous with our policy network to generate a probability distribution over actions and the critic is used to estimate the expected cumulative reward of the next metro line for reducing the training variance. Specifically, the

Algorithm 1 *Generation* (G, N, B_0)

Input: Metro network graph G, the maximum number of stations N, budget B_0

Output: A metro line $Z_T = (z_1, z_2, \cdots, z_T)$

```
Initialize the metro network state s_1
for t = 1, ..., N do
    Update the feasibility rules vector F(s_t, G)
    Update the cost b_0
    if F(s_t, G).any() and b_0 < B_0 then
        Compute embedded features L^t and V^t of the current state by the encoder
        Update hidden state h^t based on the decoder according to Equation (7)
        Generate a probability distribution over total stations by the attention mechanism according to Equation (8), Equation (9) and Equation (10)
        Select the station z_t with the probability p(z_t|s_t, G) in Equation (10)
        Updata the metro network state s_{t+1}
    else
        Terminate the metro line expansion
    end if
end for
return Solution = Z_T
```

gradient of Equation (6) according to the study [31] is

$$\nabla J(\theta|G) = \mathbb{E}_{Z\sim\pi_\theta}[(\omega(Z|G) - b(G))\nabla_\theta \log p_\theta(Z|G)], \tag{11}$$

where $\omega(Z|G)$ presents the transportation demands achieved by the metro line Z, $b(G)$ is the estimated expected cumulative reward of the next metro line, and $p_\theta(Z|G)$ is the generating probability of Z. According to Equation (11), during the interaction between the agent and the environment, if a generated metro line Z achieves more transportation demands than the current estimated expected cumulative reward $b(G)$, the policy network is trained to increase the probability of this line Z, and more transportation demands lead to larger increases. In our study, the critic is a neural network which consists of three convolutional layers and two fully-connected layers. It takes the initial state of the metro network as input, and outputs a scalar to estimate the expected cumulative reward of the next metro line. The training details are shown in Algorithm 2.

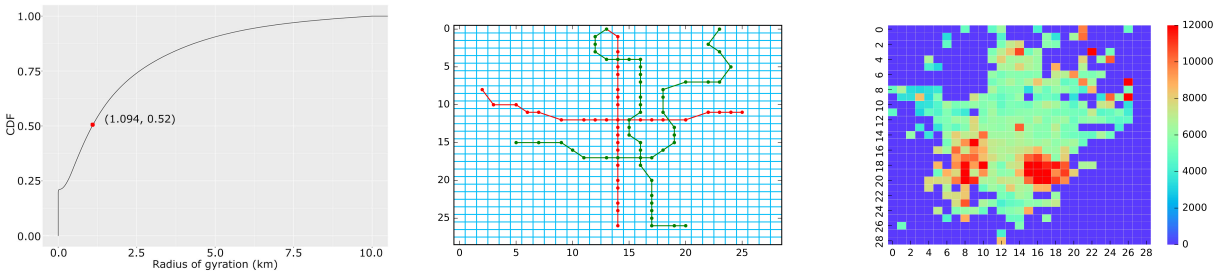
5 EXPERIMENTS

In this section, we conduct a case study to demonstrate the effectiveness of our RL method. Our codes are available online[1].

5.1 Data and Pre-processing

The case study is conducted based on the metro network in Xi'an, Shaanxi Province, China. Its first line started operation on September 16, 2011, and four lines were in operation by January 2020. Our experimental data, coming from a citywide 3G cellular data network, records the location information of 24,770,715 mobile phone users in Xi'an from October 1, 2015, to October 31, 2015. With the

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



(a) The CDF of the radius of gyration (b) Metro network (c) House price (RMB)

Figure 3: The current city operational status. (a) presents the cumulative distribution function of the radius of gyration. (b) shows the four metro lines currently operating in Xi'an, of which the two red lines were opened before our experimental data, and the two green lines were opened later. (c) presents the distribution of grid average house price in 2015, which is used to characterize the index of development of grid.

Algorithm 2 Actor-Critic Training

Require: Metro network graph G, batch size B, training epoch E, buffer R, the maximum number of stations N, budget B_0

Initialize actor parameters θ
Initialize critic parameters θ_c
Initialize buffer R
for epoch = 1, ..., E **do**
  for *instance* = 1, ..., B **do**
    $Z^i = Generation\,(G, N, B_0)$
    Calculate the satisfied transportation demands $\omega(Z^i|G)$
    Store the generated metro line Z^i and $\omega(Z^i|G)$ in R
  end for
  Calculate the $b(G)$ by critic
  Update the actor parameters using the sampled gradient:
$\nabla J(\theta|G) \leftarrow \frac{1}{B}\sum_{i=1}^{B}(\omega(Z^i|G) - b(G))\nabla_\theta \log p_\theta\left(Z^i|G\right)$
  Update critic by minimizing the loss:
$L_c \leftarrow \frac{1}{B}\sum_{i=1}^{B}(\omega(Z^i|G) - b(G))^2$
  Update the parameters of actor $\theta \leftarrow \text{Adam}(\theta, \nabla J(\theta|G))$
  Update the parameters of critic $\theta_c \leftarrow \text{Adam}(\theta_c, \nabla L_c)$
end for

help of a Hadoop cluster with 3 master nodes and 10 slave nodes[2], we process each user's original location records to a sequence of stay points that represents this user's space activity trajectory, referring to study [33]. Each stay point corresponds to a semantic location with exact latitude and longitude.

Our metro network expansion is conducted on a city represented by grids. To determine the grid size, with the above user trajectory data, we calculate each user's radius of gyration which is a metric to distinguish users' mobility patterns, and chose the distinct geographic distance which separates total users equally into two main groups as the grid size [34]. Figure 3(a) depicts the cumulative distribution function of the radius of gyration, in which users with the radius of gyration less than 1094 meters account for 50.2% of the total. Finally, we set each grid as a square with 1000 meters width and divide the study area into 29×29 grids. Correspondingly, we set the size of the filter in Appendix A as 5×5 according to study [22], which means that the distance between the adjacent stations is between 1000 meters and 2000 meters.

With the above in mind, the realistic metro network with 4 lines in Xi'an is presented in Figure 3(b), where the red lines represent the 2 lines that existed before October 2015, the green lines represent the subsequently opened 2 lines, and the dots represent stations. Mapping the above user stay points into grids, we calculate the OD trips between any two grids. Figure 3(c) presents the distribution of the average house price in 2015. The average house price of grid g_i is used to characterize its index of development D_i, which is applied to the calculation of social equity in Appendix C.

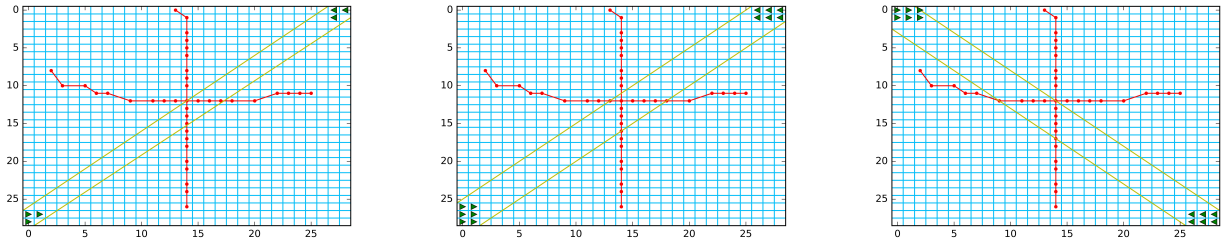
In this study, we use only the two red lines opened before the experimental data as the existing metro network, and then design the next metro line. As for the cost of construction, we set 5 billion RMB for each station, and 1 billion RMB per kilometer line referring to the study [30].

5.2 Baselines and Performance Evaluation

We compare our RL method with the following baselines. The implementation details of each method are in Appendix B.

- **Mathematical Programming Method (MP)** [30]. This method formulates the metro network expansion problem as a mathematical integer programming model. With predefined corridors and specified end stations, this method adopts a solver to obtain solutions.
- **Greedy Strategy Method (GS)** [21]. This method first selects the edge that satisfies the maximum objective, such as OD trips, and then gradually extends the current metro line by adding the surrounding stations, yielding the maximum objective.
- **Genetic Algorithm Method (GA)** [26]. This method first generates the initial population (the set of feasible metro lines), and then selects individuals (metro lines) to conduct crossover and mutation for better generations.
- **Ant Colony Algorithm (ACA)** [32]. This method first introduces a pheromone related probabilistic rule to guide the station selection. Then, according to the objective satisfied

[2]Each master node uses a dual Intel E5-2680 v4 CPU @ 2.4GHz with 14 cores. Each slave node has a dual Intel E5-2650 v4 CPU @ 2.4GHz with 12 cores. The total RAM is 1.5 TB and the total storage is 260 TB. The nodes run on CentOS release 6.8 with Hadoop 2.6.0-cdh5.5.0.



(a) The first predefined corridor (b) The second predefined corridor (c) The third predefined corridor.

Figure 4: The predefined corridors and the specified end stations. The areas between two yellow lines are the predefined corridors, and the green triangles are the specified end stations. In this experiment, the metro lines are only allowed to be built within these corridors, and the metro lines of MP need to end at the specified end stations.

by the newly generated metro lines, this method updates the pheromone of each edge for better station selection.

Existing study [12] employs the satisfied OD trips to evaluate the solution performance. Our study also uses satisfied OD trips, but also considers the social equity factor to expand the metro network as shown in Equation (3). Therefore, setting the weights α_1 and α_2 to different values, we use the following three objective functions to evaluate the performance of different methods:

$$\omega_1 = R_{od}(Z), \tag{12}$$

$$\omega_2 = 0.5 * R_{od}(Z) + 0.5 * R_{ac}(Z), \tag{13}$$

$$\omega_3 = R_{ac}(Z), \tag{14}$$

where $R_{od}(Z)$ is the added OD trips by metro line Z (refer to Equation (1)) and $R_{ac}(Z)$ is the social equity indicator of Z (refer to Equation (2)). Before metric calculation, we rescale the OD trips between grids and the house price of each grid into $[0, 1]$ by dividing them by their respective maximums.

5.3 Comparison with Baselines

5.3.1 Comparison with MP in predefined corridors. To make a comparison between the MP and our RL method, we predefine 3 corridors, and utilize both methods to build a new metro line in each corridor. Except for the corridors, we specify the end stations and enumerate all the edges that satisfy our constraints in Section 3 for MP, as these are an essential part of MP. Figure 4 presents the 3 corridors and the corresponding end stations, and Table 1 illustrates the performance. The performance of the expanded metro line varies greatly from corridor to corridor, which indicates that expert guidance seriously affects the performance of metro lines. Even within a single corridor, the poor performance suggests that specifying end stations may have a bad impact on the performance of the metro line. The results of the MP must go through the specified end stations. Due to budget limitations, the areas satisfying more transportation demands may not have the opportunity to build a metro line, which can explain this poor performance. When the budget is small, this situation is more obvious. Therefore, while the MP can get an optimal solution on a small scale, its results rely heavily on expert guidance, which lacks reliability. On the contrary, our RL method requires no expert guidance and obtains near-optimal solutions. In the next section, we will demonstrate the advantages of our RL method over the whole city space.

Table 1: Performances of MP and our RL method in predefined corridors.

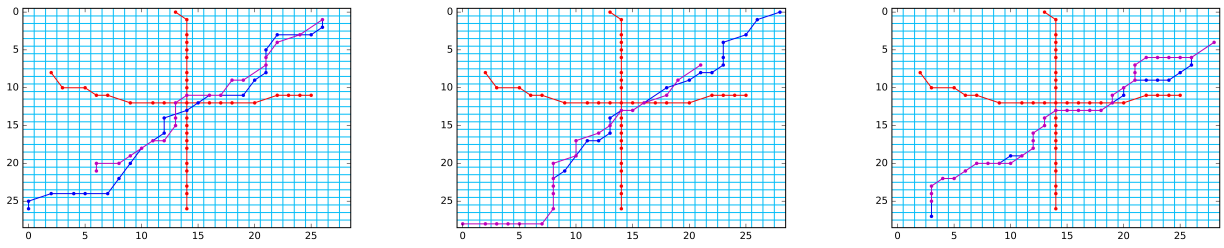
		Budget=210		Budget=270	
	Obj.	MP	RL	MP	RL
Corridor1	ω_1	41.043	**44.161**	**45.899**	44.888
	ω_2	42.53	**46.820**	50.053	**50.393**
	ω_3	44.162	**49.430**	**54.842**	54.772
Corridor2	ω_1	51.404	**54.376**	**59.160**	58.694
	ω_2	48.217	**54.406**	**60.014**	58.575
	ω_3	45.501	**55.495**	**65.189**	64.541
Corridor3	ω_1	**39.858**	39.222	**42.186**	39.623
	ω_2	38.826	**40.574**	42.240	**42.259**
	ω_3	39.724	**42.123**	43.724	**44.431**

5.3.2 Comparison with heuristic methods in the whole city space. In this section, we conduct both heuristic baselines and our RL method to design the next metro line, using the whole city rather than specified corridors. Each method is executed five times, and the average performances are shown in Table 2. Our RL method achieves significant performance improvements in all cases. The GS only focuses on the local information. It cannot consider the remote transportation demands to make a better decision, and naturally achieves a poor performance. For the GA, its heuristic operators are likely to lead to invalid solutions during the crossover and mutation, due to the shape constraints of the metro line. The low efficiency of heuristic operators hinders the evolution of better solutions. For the ACA, without the specified initial station, its strategy of randomly selecting the initial station makes no guarantee for a good performance. The statistical comparison results are shown in Appendix D.

In addition to satisfying transportation demands, the stability of the solution is also important, since a method with large solution differences cannot provide convincing results for city decision makers. According to Table 2, our RL method has a relatively steady performance over the GA and ACA (small standard deviation). To intuitively perceive the differences between the results, we further map the solutions of each method on the city grids. As shown

Table 2: Performances of heuristic methods and our RL method in the whole city space. Each entry is the final average objective ± standard deviation across 5 trials.

	Budget =210			Budget=270		
	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3
GS	19.082±0.000	26.718±0.000	24.333±0.000	19.082±0.000	28.614±0.000	35.319±0.000
GA	25.672±2.013	20.969±3.703	20.793±2.529	28.904±0.547	25.160±1.476	27.666±1.318
ACA	29.957±0.949	39.611±0.700	48.836±2.236	30.264±1.129	40.476±1.332	50.830±1.761
RL	**59.268**±0.161	**56.177**±0.320	**57.765**±0.062	**64.306**±0.582	**62.106**±0.088	**65.775**±0.166



(a) GA (b) ACA (c) Our RL method

Figure 5: The expanded metro lines achieving the maximum and minimum OD trips of different methods with a budget of 270 billion RMB across 5 trials. The blue (violet) line achieves the maximum (minimum) OD trips. Our RL method has a relatively steady performance over GA and ACA.

in Figure 5, the solutions of our RL method differ relatively little in shape, while the solutions of GA and ACA are quite different. Therefore, considering the satisfied transportation demands and the stability of the solution, our method has great advantages.

5.4 Comparison with Realistically Planned Metro Line

In this section, we explain our solutions from a realistic perspective. After using estimation to make multiple plans within a reasonable cost range, we will use the most practical one, which uses a budget of 270 billion RMB, as an example. The results are shown in Figure 6.

When considering only the OD trips, the expanded metro line starts from the Xi'an Nan Railway Station at the lower left corner, passes through the city center, and then goes in the direction of the Terracotta Army at the upper right corner, as shown in Figure 6(a). When considering OD trips and social equity as equally important, the expanded metro line is more like a partial combination of two lines planned in reality: its lower left part is similar to the realistically planned line 6, and its upper right part is similar to the line 3 currently in operation, as shown in Figure 6(b). This case shows the rationality of considering social equity in transportation planning. Figure 6(c) is the result of considering only social equity. By comparing with Figure 3(c), we find that this metro line has passed through the grids with a high development level, which intuitively demonstrates the effectiveness of our method.

Different objectives lead to different metro networks. The objectives of metro expansion vary with different cities and stages, which may require existing methods, whether heuristics or predefined corridors, to be revised. By changing the reward function, our method can be easily extended to different objectives without problem-specific knowledge. Therefore, our method is generic, and suitable for metro expansion.

5.5 Multiple Lines Expansion

Considering OD trips and social equity as equally important, we sequentially design the second metro line with our expanded line in Figure 6(b) as existing. The violet line in Figure 7 represents the expanded line. Its shape is like a partial combination of the 2 subsequently opened lines after October 2015. In this gradual way, we expand the metro network with multiple lines.

6 CONCLUSION

This paper presents a RL based method to solve the city metro network expansion problem. By formulating metro line expansion as a process of sequential station selection, we train an actor-critic model to design the next metro line. Through a case study, our method shows great advantages over baselines, achieving higher transportation demands and showing better stability, even without expert guidance. In addition, compared with the realistically planned metro lines, the effectiveness of our method is further confirmed.

REFERENCES

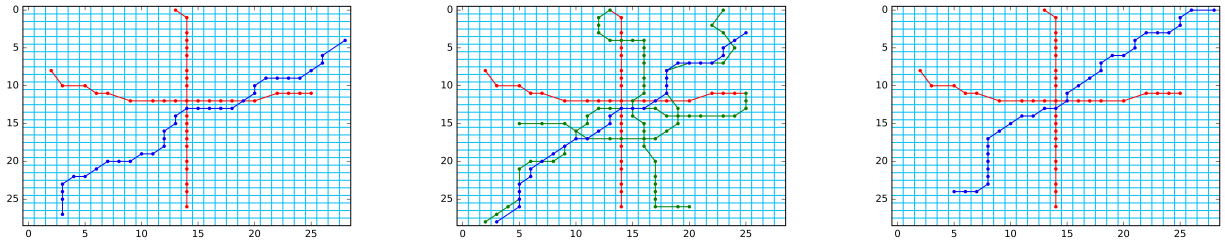
[1] Elisabete Arsenio, Karel Martens, and Floridea Di Ciommo. 2016. Sustainable urban mobility plans: Bridging climate change and equity targets? *Research in Transportation Economics* 55 (2016), 30–39.

[2] Hamid Behbahani, Sobhan Nazari, Masood Jafari Kang, and Todd Litman. 2019. A conceptual framework to formulate transportation network design problem considering social equity criteria. *Transportation Research Part A: Policy and Practice* 125 (2019), 171–183.

[3] Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. 2016. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940* (2016).

[4] Giuseppe Bruno, Michel Gendreau, and Gilbert Laporte. 2002. A heuristic for the location of a rapid transit line. *Computers & Operations Research* 29, 1 (2002), 1–12.

[5] Partha Chakroborty. 2003. Genetic algorithms for optimal urban transit network design. *Computer-Aided Civil and Infrastructure Engineering* 18, 3 (2003), 184–200.



(a) ω_1 (b) ω_2 (c) ω_3

Figure 6: The expanded metro lines under different objectives. The blue lines are our expanded results.

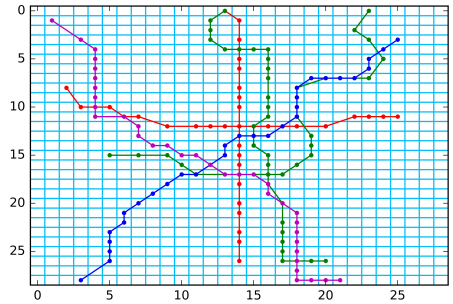


Figure 7: Multiple lines expansion. Taking into consideration OD trips and social equity, the blue and violet lines are the first and second expanded lines, respectively.

[6] Hélène Dufourd, Michel Gendreau, and Gilbert Laporte. 1996. Locating a transit line using tabu search. *Location Science* 4, 1-2 (1996), 1–19.

[7] Wei Fan and Randy B Machemehl. 2006. Using a simulated annealing algorithm to solve the transit route network design problem. *Journal of transportation engineering* 132, 2 (2006), 122–132.

[8] Reza Zanjirani Farahani, Elnaz Miandoabchi, Wai Yuen Szeto, and Hannaneh Rashidi. 2013. A review of urban transportation network design problems. *European Journal of Operational Research* 229, 2 (2013), 281–302.

[9] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

[10] Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska. 2012. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1291–1307.

[11] Gabriel Gutiérrez-Jarpa, Gilbert Laporte, and Vladimir Marianov. 2018. Corridor-based metro network design with travel flow capture. *Computers & Operations Research* 89 (2018), 58–67.

[12] Gabriel Gutiérrez-Jarpa, Carlos Obreque, Gilbert Laporte, and Vladimir Marianov. 2013. Rapid transit network design for optimal cost and origin–destination demand capture. *Computers & Operations Research* 40, 12 (2013), 3000–3009.

[13] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Konstantinos Kepaptsoglou and Matthew Karlaftis. 2009. Transit route network design problem. *Journal of transportation engineering* 135, 8 (2009), 491–505.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Vijay R Konda and John N Tsitsiklis. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*. 1008–1014.

[18] Michael Kuntz and Marco Helbich. 2014. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science* 28, 9 (2014), 1904–1921.

[19] Gilbert Laporte and Juan A Mesa. 2015. The design of rapid transit networks. In *Location science*. Springer, 581–594.

[20] Gilbert Laporte, Juan A Mesa, and Francisco A Ortega. 2000. Optimization methods for the planning of rapid transit systems. *European Journal of Operational Research* 122, 1 (2000), 1–10.

[21] Gilbert Laporte, Juan A Mesa, Francisco A Ortega, and Ignacio Sevillano. 2005. Maximizing trip coverage in the location of a single rapid transit alignment. *Annals of Operations Research* 136, 1 (2005), 49–63.

[22] Gilbert Laporte and Marta MB Pascoal. 2015. Path based algorithms for metro network design. *Computers & Operations Research* 62 (2015), 78–94.

[23] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[24] Kevin Manaugh, Madhav G Badami, and Ahmed M El-Geneidy. 2015. Integrating social equity into urban transportation planning: A critical evaluation of equity objectives and measures in transportation plans in North America. *Transport policy* 37 (2015), 167–176.

[25] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[26] Mahmoud Owais and Mostafa K Osman. 2018. Complete hierarchical multi-objective genetic algorithm for transit network design problem. *Expert Systems with Applications* 114 (2018), 143–154.

[27] Yanshuo Sun, Paul Schonfeld, and Qianwen Guo. 2018. Optimal extension of rail transit lines. *International Journal of Sustainable Transportation* 12, 10 (2018), 753–769.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[29] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*. 2692–2700.

[30] Yi Wei, Jian Gang Jin, Jingfeng Yang, and Linjun Lu. 2019. Strategic network expansion of urban rapid transit systems: A bi-objective programming model. *Computer-Aided Civil and Infrastructure Engineering* 34, 5 (2019), 431–443.

[31] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

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

[33] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. 2009. Mining individual life pattern based on location history. In *2009 tenth international conference on mobile data management: systems, services and middleware*. IEEE, 1–10.

[34] Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. 2017. Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science* 31, 7 (2017), 1293–1313.

[35] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1040–1048.

ACKNOWLEDGMENTS

The corresponding author Xi Zhao is a Tang Scholar. This work was supported by the National Natural Science Foundation of China (Grant No. 91746111).

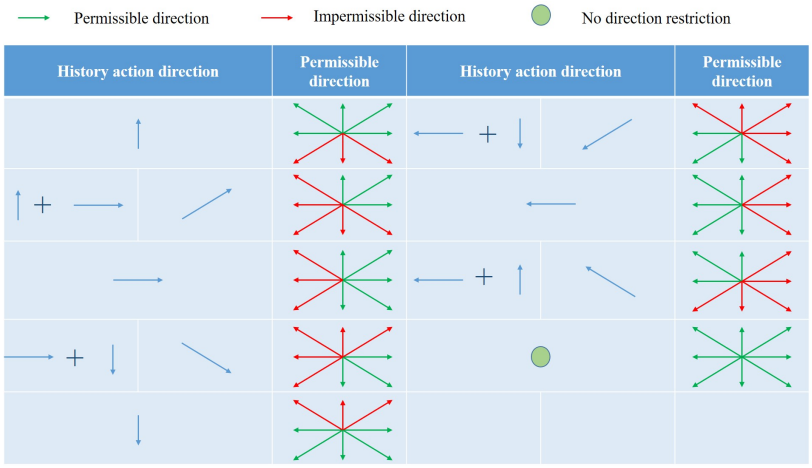


Figure 8: Action direction rules.

Table 3: Bootstrap mean and 95% confidence bounds for the experiments. 10K bootstrap iterations and the pivotal method are used.

	Budget =210			Budget =270		
	ω_1	ω_2	ω_3	ω_1	ω_2	ω_3
GS	19.082 (19.082,19.082)	26.718 (26.718,26.718)	24.333 (24.333,24.333)	19.082 (19.082,19.082)	28.614 (28.614,28.614)	35.319 (35.319,35.319)
GA	25.672 (23.985,27.164)	20.969 (17.720,23.410)	20.793 (18.568,22.552)	28.904 (28.505, 29.335)	25.160 (24.017, 26.288)	27.666 (26.852,28.843)
ACA	29.957 (29.273,30.778)	39.611 (39.093,40.135)	48.836 (47.286,50.719)	30.264 (29.396, 31.118)	40.476 (39.384, 41.474)	50.830 (49.346, 52.143)
RL	**59.268** (59.143,59.392)	**56.177** (55.917, 56.408)	**57.765** (57.718,57.811)	**64.306** (63.755, 64.762)	**62.106** (62.035, 62.171)	**65.775** (65.667, 65.910)

A FEASIBILITY RULES

In this section, we present the feasibility rules $F(s_t, G)$ in Equation (4), which constrains the agent's action, to ensure that the expanded metro line satisfies the constraints in Section 3. For the first constraint, we use a certain filter with the current selected station in the center to limit the selecting range of optional stations for the agent. The filter is a square shape with $m \times m$ grids, and the agent can only select the next station within this filter. For the second constraint, in terms of shape assurance, we design the action direction rules based on historical actions to determine the optional action in the next step as shown in Figure 8. Except for this action direction rules, the existing metro network also affects the action selection. During the expansion, the expanded metro line is allowed to connect with existing lines to form transfer stations, but it cannot coincide with the existing lines. To achieve the above, we present the feasibility rules $F(s_t, G)$ as a binary vector with dimension n^2, where the element is set to 1 if this station is optional; otherwise the element is set to 0.

In addition to the feasibility rules $F(s_t, G)$, for the third constraint, once the number of selected stations reaches the upper limit N, the expansion process is terminated. For the fourth constraint, per unit length of the metro line and per station consume a certain cost. During the interaction between the agent and the environment, once the cost of construction exceeds the budget B_0, the expansion process is terminated.

With these above, the agent interacts with the environment as follows. Initially, all of the elements of the vector $F(s_t, G)$ are 1, and the agent selects a station based on the parameterized policy π_θ. Then, the $F(s_t, G)$ is updated according to the filter and the action direction rules. From this time on, the agent is only allowed to select a station with the element as 1 in the $F(s_t, G)$ according to the policy π_θ. This process will be repeated until all the elements of the vector $F(s_t, G)$ are 0, or the number of selected stations reaches the upper limit N, or the cost of construction exceeds the budget B. An example is shown in Figure 9.

B METHOD DETAILS

This section describes the implementation details of our RL method and the compared methods in Section 5.2. All the experiments are conducted on a desktop with an E5-2680 v4 @ 2.40GHz CPU and a TITAN Xp GPU.

For our RL method, we set the two 1-dimensional convolutional neural networks in encoder with 128 filters, and the LSTM in decoder with a state size of 128. The parameters in our network are

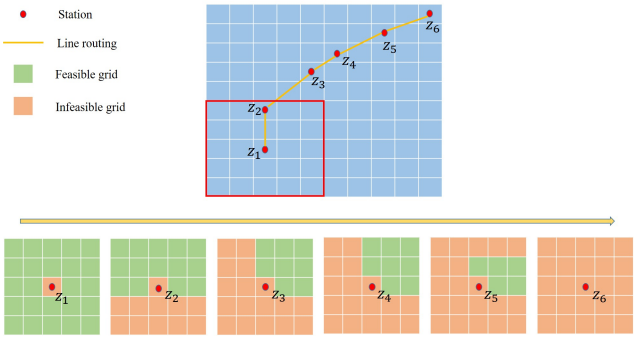


Figure 9: An example of metro line expansion.

initialized with Xavier initialization [9]. We adopt the Adam optimizer [16] with learning rate 10^{-4} to train our network. During the training, the batch size B is 128, and the training epoch E is 3500. After training, we employ the parameters which achieve the maximum objective in the training process to generate the metro line.

For the MP, we predefine the corridor and end stations, and employ Gurobi to obtain the new metro line. The newly constructed metro line must stay in the predefined corridor and end at the specified end stations.

For the GA, its fitness function is the objective in Equation (3). After our parameter tuning, we set the initial population size as 500, and the maximum number of iterations as 3500. The crossover is conducted at the common station of two selected metro lines, and each station on a metro line is allowed to mutate. After crossover and mutation, the original metro lines that fail to evolve and the new metro lines that satisfy our constraints in Section 3 are preserved. Considering the low efficiency of heuristic operators, we set both the crossover probability and mutation probability as 0.9. At last, the metro line with the maximum fitness during the evolution process is output as the final solution.

For the ACA, we set the maximum number of iterations as 3500, with each iteration containing 128 instances. All the other parameters are the same as study [32]. In each iteration, the initial station for each instance is randomly selected, and then the subsequent station selection is guided by the probabilistic rule. At last, the metro line with the maximum objective during the evolution process is output as the final solution.

In the phase of mapping the solution on the city grids, the final solutions of all methods are modified using gurobi to be more realistic.

C SOCIAL EQUITY

As an important transportation system, the metro network benefits the areas traversed by it for convenient access to other areas, which may be reflected in people's access to education, economic activity, and other aspects. Referring to [2], we consider an accessibility variable as each area's distributable benefit obtaining from the metro network, and adopt a utilitarianism theory to aggregate each area's distributable benefit to measure the social equity.

For an area i, its accessibility index Ac_i is calculated as:

$$Ac_i = \sum_j D_j F\left(c_{ij}\right), \tag{15}$$

where c_{ij} is the generalized travel cost between area i and area j, $F\left(c_{ij}\right)$ represents the function of resistance against traveling between area i and area j, and D_j represents the compound index of development of area j. Specifically, $F\left(c_{ij}\right)$ is defined as:

$$F\left(c_{ij}\right) = F\left(t_{ij}\right) = e^{-\beta t_{ij}}, \tag{16}$$

where t_{ij} is the travel time between area i and area j, and β is an adjustment parameter. D_j is defined as:

$$D_j = \sum_k w_k d_{j,k}, \tag{17}$$

where $d_{j,k}$ is an economic variable of type k for area j and w_k is the weight. In our study, we consider that t_{ij} is proportional to the distance between i and j, and take the house price of area j as its index of development D_j [18].

Then, under a utilitarianism theory, the social equity indicator R_{ac} in the transportation network planning is defined as:

$$R_{ac} = \sum_i Ac_i, \tag{18}$$

R_{ac} measures the total benefits the transportation network brings to society. We prefer the metro network to achieve greater social equity indicator R_{ac}.

D STATISTICAL COMPARISON OF DIFFERENT METHODS' PERFORMANCES

In order to make our experimental results more convincing, we employ the bootstrap and significance testing to evaluate the methods' performances in Section 5.3.2, referring to study [13]. Table 3 shows the bootstrap mean and 95% confidence bounds on our experiments, which demonstrates that our method performs best in a statistical sense. In fact, we have conducted power analysis for the choice of the sample size, and the analysis demonstrates that 5 trials in our experiments are enough.

We further conduct a difference test to compare the performances of different methods, referring to study [13]. Limited by space, we only present the comparison of our RL method with ACA under a budget of 270 billion RMB and the same objective ω_1. Assuming the null hypothesis H_0, that there is no difference in performance between these two methods, the bootstrap confidence interval test estimates the confidence interval for the difference between the mean performances of these two methods as [33.1925, 34.927] at significance level 0.05. The confidence interval does not include 0 and both bounds are positive. Thus the null hypothesis H_0 is rejected, and we can be confident at 95% that the performance of our RL method is superior to that of the ACA method.