Finding Feasible Routes with Reinforcement Learning Using Macro-Level Traffic Measurements

Mustafa Can Ozkan[1]
SpaceTimeLab, University College London, UK

Tao Cheng
SpaceTimeLab, University College London, UK

Abstract

The quest for identifying feasible routes holds immense significance in the realm of transportation, spanning a diverse range of applications, from logistics and emergency systems to taxis and public transport services. This research area offers multifaceted benefits, including optimising traffic management, maximising traffic flow, and reducing carbon emissions and fuel consumption. Extensive studies have been conducted to address this critical issue, with a primary focus on finding the shortest paths, while some of them incorporate various traffic conditions such as waiting times at traffic lights and traffic speeds on road segments. In this study, we direct our attention towards historical data sets that encapsulate individuals' route preferences, assuming they encompass all traffic conditions, real-time decisions and topological features. We acknowledge that the prevailing preferences during the recorded period serve as a guide for feasible routes. The study's noteworthy contribution lies in our departure from analysing individual preferences and trajectory information, instead focusing solely on macro-level measurements of each road segment, such as traffic flow or traffic speed. These types of macro-level measurements are easier to collect compared to individual data sets. We propose an algorithm based on Q-learning, employing traffic measurements within a road network as positive attractive rewards for an agent. In short, observations from macro-level decisions will help us to determine optimal routes between any two points. Preliminary results demonstrate the agent's ability to accurately identify the most feasible routes within a short training period.

2012 ACM Subject Classification Computing methodologies → Q-learning

Keywords and phrases routing, reinforcement learning, q-learning, data mining, macro-level patterns

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.58

Category Short Paper

1 Introduction

The topic of finding routes between two points has been studied in many different fields, such as computer systems, transportation systems and communication networks. The majority of research concentrates on route optimisation, seeking to reduce travel time or distance or to maximise operational efficiencies, such as the maximum number of taxi customers or the maximum storage of a delivery truck. These studies, which employ mathematical optimisation techniques, include optimisation constraints such as the truck's maximum cargo capacity and minimise/maximise the objective function of the main aim, such as travel time. They often take into account the average travel time on a route depending on the length of the road, the timing of the traffic lights, or occasionally the traffic situation, including actual or historical traffic flow and speeds. They also factor in user preferences from surveys or GPS

[1] corresponding author

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 58; pp. 58:1–58:6
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

to generate the most popular routes ahead of time. All of these aspects make optimal route research hard and costly when using multiple data sources. As more realistic findings are sought, models and algorithms become increasingly complex and computationally expensive to investigate every aspect of the traffic situation and road network infrastructure.

In this study, we assume that all of these factors, including traffic user preferences, traffic conditions, and road network features, are already represented in macro-level historical observations. We attempt to extract the most feasible routes using macro-level measurements, in other words, by using the most popular road segments in a road network. We aim to train a reinforcement learning agent to mimic human behaviours for route choices and use the agent to detect the most taken routes between any two nodes which might or might not be optimal routes at that time period under certain traffic conditions.

Although the approach does not guarantee route optimisation, it will identify the most practical and feasible route options at that time from the reflection of the preferences of mass mobility actions. The relevant studies on the algorithms and RL-related studies in route finding in the transport research sector will be briefly discussed in the next section.

2 Related Studies

The principles of routing in transportation are based on the most well-known problems including the travelling salesman problem (TSP), the vehicle routing problem(VRP), and the shortest path problem. They are the primary subjects with the goal of determining the best transport strategies over a road network from a source to a destination. The classic Dijkstra algorithm from 1959 [3] is where the history of discovering shortest paths in networks begins. Heuristic algorithms such as A*[5] concerned with the heading to the destination were created because the Dijkstra algorithm has a vast solution space. These are the core algorithms for static networks, and they only produce one shortest path. However, due to the dynamic nature of road networks, various algorithms for problems involving short paths are introduced with dynamic variables.

Reinforcement learning algorithms have been combined with these conventional techniques to tackle common routing problems such as VRP and TSP [9][14]. Recent RL research has begun to focus on applying deep learning techniques to TSP [13] and VRP problems [1]. These studies are not only limited by classical problems but also attempt to solve optimisation problems in shared transportation [11] and taxi systems [8] by considering future demands. Basically, they define reward systems for desired outcomes such as potential high-demanded areas for taxis. Some studies[2][10] introduce dynamic variables such as energy consumption, and customer request to design some optimisation constraints. This also affects routing studies for passengers[4].

All these studies focus on only the main goal to define a reward system. Our approach will employ mid-rewards to encourage the agent to mimic human behaviours based on observations to find feasible routes. According to the studies [4][7][6], classical routing algorithms are not the best methods to find feasible routes in large systems because of the time complexity and insufficient capabilities of considering only distance costs. On the other hand, All of the aforementioned studies focus on finding the best options within predefined rules, assumptions and constraints. We aim to remove all the pre-defined assumptions and constraints.

Listing 1 Pseudo code for the Q-Learning.

```
Input: Macro-Level measurements (traffic flow), Graph representation
Output: Q-Table
Initialise Reward R(s,a) and Q-table (Q(s,a))
for i:0 to the number of iteration
    Select a random node and its neighbours
    Update the Q-value of the node pairs with the equation
Return Updated Q-table
end
- Select Feasible Route: Reach destinations by selecting the highest
q-values from the starting state
- Derivate other routes: Let the agent choose other best options
```

3 Methodology

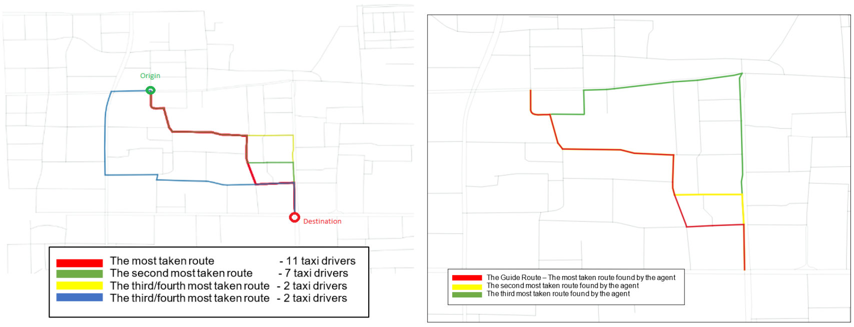
The study's core part is based on a reinforcement learning algorithm called Q-learning. It is a branch of machine learning where an agent maximises its cumulative reward by collecting rewards based on its actions and interactions with an environment. The environment can be modelled mathematically or can be model-free by only focusing on certain rewards that encourage behaviours.

Our approach uses model-free Q-Learning algorithm taking traffic flow values on road segments as the reward. The purpose of the approach is to extract significant routing behaviours from macroscopic observations. It is an offline approach, which implies the agent tries to mimic given behaviours using historical observations without having any impact on the environment. A directed graph is used to represent the environment that represents road networks. Road segments are the edges and intersections are the nodes in this graph. At each intersection, the agent can choose the next road segment (state) to travel through by considering the normalised flow values, in other words, the rewards. These mid-rewards encourage the agent to choose the most taken routes between any two nodes. To help the agent reach the destination point, the highest point is given to the destination points.

These rewards in Q-Learning have an impact on the equation that modifies Q-values in a table (Q-table) displaying the values computed based on state and action pairs. In our approach, we used the Bellmann equation, which performs computations based on states, current and projected rewards, and current Q-values. The Bellman equation;

$$Q_{new}(s_t, a_t) = (1 - a) * Q(s_t, a_t) + \alpha * (R(s, a) + \lambda * maxQ(s^{'}, a^{'}))$$

Where R is the reward value at state s in the taken action a. The discount variable λ controls the rate at which future rewards will affect the Q-values. The learning rate, or α determines how the current state and actions will influence the Q-values. With the Bellman equation, all the Q-values can be updated in the Q-table in each interaction. This is essential in the agent's training stage. After the training is completed, any starting point can be selected as a current state and the agent can choose the best q-value chain reaching the destination point represented by another state. The highest q-values at each state are consecutively selected to complete this process. The total Q-value value is not required to have the highest value. There might be other route options with higher q-values in total but they are not the best options showing significant mobility patterns. A randomness parameter is introduced to allow the agent to select q-values other than the best one in order


Origin
Destination
The most taken route - 11 taxi drivers
The second most taken route - 7 taxi drivers
The third/fourth most taken route - 2 taxi drivers
The third/fourth most taken route - 2 taxi drivers
The Guide Route – The most taken route found by the agent
The second most taken route found by the agent
The third most taken route found by the agent

(a) Truth routes taken by taxi drivers. **(b)** Routes found by the RL agent.

Figure 1 Routes between selected origin-destination pairs.

to derive additional route alternatives. By comparing the distances between the users' actual routes and the routes discovered by the approach, path similarity algorithms can validate the method.

4 Case study

To test the proposed Q-learning approach, a well-known taxi trajectory data set collected in Beijing by Microsoft[12] is used. The OSMnx package, which uses OpenStreetMap as a source, is used to extract the road network. For our approach, there is no need to use micro-level or individual-level measurements such as GPS points or trajectories. However, this data set is helpful to demonstrate the effectiveness of the approach by providing taxi trip trajectories to be used in the validation. We used map-matching algorithms to aggregate all the individual trips for one day in the study area in order to obtain the traffic flow values that the algorithm needs as input. So, we can have the number of taxis at each road segment throughout the entire network.

The dataset contains the trajectory data for nearly 10,000 taxis for the days between February 2nd and February 8th in Beijing. We selected the first-day data of 1000 trips in Beijing's central region for simplicity and due to computation costs. For the study area, there are 657 nodes representing intersections and 1542 edges between these nodes representing each road segment. To determine the origin and destination points, we selected two random point pairs in the concentrated areas having the highest flow values by observing the dataset. Although it is not guaranteed that all trips begin and end at these points, they are sufficient to test our methodology and compare the feasible routes with the real routes passing between these two points. This step is only performed for validation aims.

4.1 Q-learning and Initial Results

During the training phase, the discount and learning rate are chosen as 0.8 and the Q-value table is updated one million iterations. For the study area, this procedure takes 150 seconds to complete.

We detected 31 taxi drivers and 13 different route options taken by these taxi drivers in real life between the origin and destination points in figure 1a. For the visualisation, we showed only the 4 most taken routes by taxi drivers in figure 1a and the three most feasible routes found by the RL agent in figure 1b. The best feasible route extracted from taxi drivers' behaviours by the agent has total Q-values of 6625. In this route, the agent chooses the best actions at each state until it reaches the destination point. Once the derivation process has been completed, all feasible routes can be extracted by using only traffic flow values, revealing the majority of taxi drivers' choices on their routes. Given the connectivity between nodes, it is clear that there are a finite number of physically feasible routes. The route found by the agent with the best actions is considered the most feasible route and reflects the most seen behaviour of taxi drivers. The other derived options are ordered by the distance of the total q-values from the best feasible option. These q-values can be larger or smaller than the best one in total. All these feasible routes show the preferences of the taxi drivers between these two nodes.

There are two issues with this approach. The agent can choose longer routes to collect more points or it only follows main roads with a high number of traffic flow. Therefore, the reward at the destination point should be decided carefully to encourage the agent to reach the destination point with fewer steps while also avoiding finding only the shortest paths. It should be emphasised that during the training step, all reward values derived from traffic flows are normalised.

The approach demonstrated that the agent can be trained by using only historical macro-level measurements such as traffic flow. These measurements can also be additional traffic state indicators such as traffic speed and travel time. The approach eliminates the requirement for using individual data, which are difficult to obtain and troublesome because of privacy issues. Also, individual preferences and pre-defined assumptions on behaviours can be bypassed by focusing only on macro-level patterns. The approach merely assumes that the best possible sequence of q-values will serve as a guide route.

5 Conclusion

The proposed approach uses a Q-learning-based algorithm to identify feasible route possibilities using just macro-level measurements. It does not have any assumption on the route selection and only mines the historical patterns to detect attracted route segments and extract routes between any two points from observations. This can be seen as a transaction to proceed from macro-level measurements to micro-level discoveries. Comparatively speaking, macro-level measurements are simpler and cost-friendly in terms of collecting data than individual data sets like GPS and surveys. They can be received by using sensors, cameras, or even manual counting.

The initial results show that an agent can be trained to extract feasible routes by only exploring the number of vehicles on road segments and sorting the potential options by their selection probabilities. It focuses on the attractiveness and popularity of road segments to take action for the next states. This concept will help us to develop a technique to understand route choice behaviours from macro-level patterns for future research. The approach can combine route set generation and route selection processes in route choice modelling by considering trends at any traffic state. Additionally, these feasible routes may or may not be the shortest ones, the fastest ones, or the ones that taxi drivers select because of the scenic vistas. To uncover the motivations behind these decisions, we will be conducting a more comprehensive analysis.

References

1 T Ahamed, B Zou, N P Farazi, and T Tulabandhula. Deep Reinforcement Learning for Crowdsourced Urban Delivery. *Transportation Research Part B: Methodological*, 152:227–257, 2021. doi:10.1016/j.trb.2021.08.015.

2 R Basso, B Kulcsár, I Sanchez-Diaz, and X Qu. Dynamic stochastic electric vehicle routing with safe reinforcement learning. *Transportation Research Part E: Logistics and Transportation Review*, 157, 2022. doi:10.1016/j.tre.2021.102496.

3 Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

4 Y Geng, E Liu, R Wang, Y Liu, W Rao, S Feng, Z Dong, Z Fu, and Y Chen. Deep Reinforcement Learning Based Dynamic Route Planning for Minimizing Travel Time. In *2021 IEEE International Conference on Communications Workshops, ICC Workshops 2021*, Shanghai, China, 2021. doi:10.1109/ICCWorkshops50388.2021.9473555.

5 Peter Hart, Nils Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi:10.1109/tssc.1968.300136.

6 Y Hu, L Yang, and Y Lou. Path Planning with Q-Learning. In *2021 2nd International Conference on Internet of Things, Artificial Intelligence and Mechanical Automation, IoTAIMA 2021*, volume 1948, North Carolina State University, Raleigh, NC 27695, United States, 2021. IOP Publishing Ltd. doi:10.1088/1742-6596/1948/1/012038.

7 F Jamshidi, L Zhang, and F Nezhadalinaei. Autonomous Driving Systems: Developing an Approach based on A* and Double Q-Learning. In *7th International Conference on Web Research, ICWR 2021*, pages 82–85, East China Normal University, Moe International Joint Lab of Trustworthy Software, Shanghai, China, 2021. doi:10.1109/ICWR51868.2021.9443139.

8 E Liang, K Wen, W H K Lam, A Sumalee, and R Zhong. An Integrated Reinforcement Learning and Centralized Programming Approach for Online Taxi Dispatching. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. doi:10.1109/TNNLS.2021.3060187.

9 T S Mostafa and H Talaat. An Intelligent Geographical Information System for Vehicle Routing (IGIS-VR): A modeling framework. In *13th International IEEE Conference on Intelligent Transportation Systems, ITSC 2010*, pages 801–805, Intelligent Transportation Systems Program, Nile University, 2010. doi:10.1109/ITSC.2010.5625095.

10 P Tong, Y Yan, D Wang, and X Qu. Optimal route design of electric transit networks considering travel reliability. *Computer-Aided Civil and Infrastructure Engineering*, 36(10):1229–1248, 2021. doi:10.1111/mice.12678.

11 C Wei, Y Wang, X Yan, and C Shao. Look-Ahead Insertion Policy for a Shared-Taxi System Based on Reinforcement Learning. *IEEE Access*, 6:5716–5726, 2017. doi:10.1109/ACCESS.2017.2769666.

12 Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 316–324, New York, NY, USA, 2011. Association for Computing Machinery. doi:10.1145/2020408.2020462.

13 Y Zhang, R Bai, R Qu, C Tu, and J Jin. A deep reinforcement learning based hyper-heuristic for combinatorial optimisation with uncertainties. *European Journal of Operational Research*, 2021. doi:10.1016/j.ejor.2021.10.032.

14 M Zolfpour-Arokhlo, A Selamat, S Z Mohd Hashim, and H Afkhami. Modeling of route planning system based on Q value-based dynamic programming with multi-agent reinforcement learning algorithms. *Engineering Applications of Artificial Intelligence*, 29:163–177, 2014. doi:10.1016/j.engappai.2014.01.001.