Avaliação do desempenho do aprendizado por reforço em simulação microscópica de tráfego

Liza Lunardi Lemos, Ana Bazzan, Gabriel de O. Ramos

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{lllemos, bazzan, goramos}@inf.ufrgs.br

Resumo. *Esse trabalho investiga uma maneira de diminuir o tempo médio de viagem dos veículos. Para isso, uma abordagem baseada em aprendizado por reforço para a escolha de rotas dos veículos foi utilizada. Cada veículo recebe um conjunto de rotas pré-computadas e antes de cada viagem o veículo deve escolher entre essas rotas. Para validar a abordagem foi utilizada uma rede em grade 3 x 3 e um modelo de simulação microscópico. Comparou-se o tempo médio de viagem global e por subpopulações. Os resultados foram comparados com diferentes configurações do algoritmo e com o método de alocação de tráfego DUA. Em todos os casos, nossa abordagem superou o DUA em termos de tempo médio de viagem na rede.*

Abstract. *This paper investigates ways to reduce the average travel time of vehicles. We develop a reinforcement learning method to address the route choice problem. Each vehicle receives a set of pre-computed routes and, before each trip it must choose one among these routes. We validate our approach using a 3 x 3 grid network and a microscope simulation model. It was used to compare the results the global average travel time and by subpopulations. We analyse perfomance in terms of average travel time both at global level as well as in a subpopulation level. Our method is evaluated using different configurations and is compared against the traffic assignment method called DUA. In all tested cases, our approach overperforms DUA in terms of average travel time.*

1. Introdução

Em geral, o aumento da quantidade de veículos utilizando uma rede de transporte não é proporcional ao aumento da capacidade desta infraestrutura. Soluções que expandam a infraestrutura da rede podem ser caras, além de possuir problemas ambientais. Portanto, precisa-se de uma maneira de melhorar o fluxo da rede sem mudar sua infraestrutura, ou seja, encontrar a melhor forma de utilizá-la. Os problemas nos deslocamentos enfrentados hoje nessas cidades são um problema complexo. Para resolver isso, diferentes técnicas são testadas para melhorar o fluxo na rede de transporte. Uma forma é fazer com que motoristas utilizem vias alternativas ao invés de escolher apenas rotas de menor custo para se deslocar.

Para fazer isso, podemos projetar o problema utilizando sistemas multiagente. De tal forma, as entidades da rede de transporte são modeladas como agentes. No caso deste trabalho, os veículos são modelados como agentes e a rede de transporte como ambiente. Para realizar a escolha de rotas dos veículos, uma alternativa seria utilizar o aprendizado

por reforço, no qual os agentes escolhem ações com base no conhecimento adquirido através da interação com o ambiente.

Este trabalho tem como objetivo investigar uma forma de minimizar o tempo de viagem dos usuários da rede. Para tanto, foi proposto uma abordagem baseada em aprendizado por reforço, na qual os agentes aprendem individualmente a usar a rota que melhor distribui o fluxo da rede. Cada agente possui um conjunto de rotas pré-computado. Além disso, os agentes utilizam apenas o conhecimento adquirido através de suas interações com o ambiente.

Para validar a abordagem, foi utilizada uma adaptação do cenário em forma de grade 3 x 3 proposto por [Mannion et al. 2016]. Os experimentos foram realizados no simulador microscópico SUMO (*Simulation of Urban Mobility*) [Behrisch et al. 2011]. Os resultados obtidos foram comparados ao método DUA[1] que é um módulo de alocação de tráfego do SUMO. Foram realizados experimentos com diferentes tipos de demanda e o tempo médio de viagem foi computado a nível global e separado por subpopulações. Em todos os casos, nosso método superou o DUA em termos de tempo médio de viagem na rede. Mesmo em demandas maiores, a nossa abordagem se mostrou 5% melhor que o DUA.

O texto está organizado como segue. A Seção 2 revisa os conceitos básicos sobre sistemas de transporte e aprendizado por reforço. A Seção 3 apresenta os trabalhos relacionados. Na Seção 4 são descritos o cenário e a abordagem. Na Seção 5 são apresentados os resultados e é feita uma análise sobre eles. Por fim, a Seção 6 apresenta considerações finais e trabalhos futuros.

2. Fundamentação teórica

2.1. Sistema de transporte

Nesta seção será introduzido os conceitos básicos em relação ao sistema de transporte. Para uma visão mais detalhada, recomendamos a leitura de [Bazzan and Klügl 2013] e [Ortúzar and Willumsen 2011].

Um sistema de transporte pode ser considerado um composto de duas partes: oferta e demanda. A oferta representa a infraestrutura existente (rede de transporte). Uma rede de transporte é um grafo direcionado com um conjunto de vértices e um conjunto de arcos. Os vértices representam as interseções ou cruzamentos e os arcos entre esses vértices representam as vias. Cada arco possui uma capacidade e suporta um fluxo. A capacidade é entendida como o número de unidades de tráfego que um arco suporta em um instante do tempo.

Já a demanda representa os usuários do sistema de transporte. Esses usuários se deslocam de um local a outro conforme sua necessidade. Um subconjunto de vértices contém as origens da rede de transporte, onde os veículos começam suas viagens. Outro subconjunto de vértices contém os destinos da rede de transporte, onde os veículos terminam suas viagens. Um veículo faz uma viagem de um vértice origem a um vértice destino (chamado de par OD). A viagem é feita através de uma rota, que consiste em um conjunto de vértices entre a origem e o destino.

[1] http://sumo.dlr.de/wiki/Demand/Dynamic_User_Assignment

A escolha de rotas é uma forma de conectar oferta e demanda. No entanto, se todos os veículos se deslocam pela mesma rota, a rota fica congestionada. Por outro lado, os veículos possuem outras rotas para se deslocar que podem conter menos veículos. Dessa maneira, estuda-se formas de encontrar o equilíbrio. O equilíbrio mencionado é o *user equilibrium* que representa uma condição do tráfego. Como é afirmado no primeiro princípio de Wardrop [Wardrop 1952], sob condição de equilíbrio o tráfego se organiza em redes congestionadas de modo que todas as rotas utilizadas entre um par OD tenham custos iguais e mínimos.

Para realizar experimentos nesta área existem ferramentas de simulação e diferentes níveis de abstração. Para este trabalho foi escolhido o modelo de simulação microscópico, pois com ele as entidades da rede e suas relações são apresentadas com um alto nível de detalhamento. Geralmente é mais complexo e possui desenvolvimento custoso.

2.2. Aprendizado por reforço

No aprendizado por reforço, o agente aprende um comportamento através da sua interação com o ambiente [Sutton and Barto 1998]. Para cada ação realizada no ambiente, o agente recebe um valor chamado recompensa. Dessa forma, o agente deve aprender a política que retorna recompensas melhores.

Podemos modelar o aprendizado por reforço como um processo de decisão de Markov (PDM). Um PDM é uma tupla (S, A, T, R), onde S é um conjunto de estados; A é um conjunto de ações; T é a função de transição, que modela a probabilidade de o sistema passar de um estado $s \in S$ para um estado $s' \in S$, dado que ele executou a ação $a \in A$; e R é a função de recompensa por realizar a ação $a \in A$ quando se está no estado $s \in S$.

O algoritmo escolhido para aplicar o aprendizado por reforço foi o *Q-learning*, por ser um algoritmo muito usado. O *Q-learning* possui um valor Q para cada par estado-ação, que representa uma estimativa de recompensa por executar a ação a no estado s. Uma recompensa é recebida após cada ação executada. A atualização dos valores $Q(s, a)$ é feita através da Equação (1), onde $\alpha \in [0, 1]$ é a taxa de aprendizagem (ou seja, o quanto a nova experiência influencia o valor já aprendido até o momento) e o $\gamma \in [0, 1]$ é o fator de desconto (que desvaloriza o reforço em função da diferença temporal)

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma max(Q(s', a'))) \quad (1)$$

Para o problema da escolha de rotas abordado neste trabalho, podemos modelar o PDM através de um único estado. Esta modelagem é equivalente a possuir um estado inicial com todas as ações levando a um estado final com ações de recompensa zero. Além do mais, quando um agente escolhe uma ação $a \in A$ no estado inicial, ele executa a ação desejada com probabilidade 1. Sendo assim, o valor Q será atualizado utilizando a recompensa recebida e considerando uma recompensa zero das ações do estado final. Consequentemente, a recompensa das ações futuras é irrelevante.

O conjunto de ações que o agente pode escolher é pré-definido pelo algoritmo k *Shortest Loopless Paths* (KSP) [Yen 1971], o qual encontra os k caminhos mais curtos,

desconsiderando ciclos, entre um nodo origem e um nodo destino em um grafo com arestas não negativas. Esses caminhos mais curtos são calculados numa situação em que a rede está sem congestionamento.

Para escolha da ação é utilizada a estratégia de exploração ε-gulosa, na qual a ação com melhor valor Q é escolhida com uma probabilidade de 1 - ε e uma ação aleatória é selecionada com uma probabilidade ε. Um complemento frequentemente utilizado na estratégia ε-gulosa corresponde à aplicação de uma taxa de decaimento dr ao ε, decrescendo o valor deste a cada episódio. Com isso, no início da execução há bastante exploração e, com o decorrer do tempo, há um maior aproveitamento das melhores ações.

3. Trabalhos relacionados

Esta seção apresenta alguns trabalhos que também propõem uma forma de escolha de rotas para melhorar o fluxo da rede. São discutidas as contribuições, similaridades com o presente trabalho e também suas limitações.

O trabalho de [Dias et al. 2014], utiliza o *Inverted Ant Colony Optimization* (IACO), que é baseado no algoritmo da colônia de formigas. A diferença é que ao invés de usar o feromônio para atrair as formigas, ele inverte esse efeito, repelindo-as. O sistema é definido de forma descentralizada de modo que as formigas (veículos) depositam seus feromônios nas vias que estão utilizando. Dessa forma, as formigas são repelidas de vias congestionadas, contribuindo para uma melhor distribuição do tráfego na rede. Entretanto, o feromônio precisa ser guardado por uma central, descaracterizando a modelagem descentralizada.

Outra abordagem relacionada é descrita em [Dia and Panwai 2014]. Nesta abordagem redes neurais são utilizadas para prever a conformidade dos motoristas e suas escolhas de rotas sob a influência de mensagens com informações sobre o tráfego. Entretanto, a validação é realizada em um cenário mais simples com somente três rotas. Além disso, ele analisa mais a parte da influência da mensagem nos agentes do que o tráfego em si.

Uma outra abordagem para o problema de escolha de rotas, só que desta vez utilizando teoria de jogos é [Galib and Moser 2011]. Nesse caso, aplica-se jogos evolutivos para alcançar um aproveitamento equilibrado da rede. A abordagem consiste em utilizar apenas experiências passadas para a escolha de rotas dos veículos, sendo que a escolha em si é feita a cada interseção da rede. No entanto, ele assume que a informação histórica está disponível para todos os veículos da rede.

Uma abordagem utilizando aprendizado por reforço pode ser vista em [Ramos and Grunitzki 2015], na qual cada agente é modelado como um *learning automata* [Narendra and Thathachar 1989]. Os agentes recebem um conjunto de rotas pré-computadas que podem ser re-calculadas ao longo da simulação com uma dada probabilidade. Um trabalho semelhante é apresentado por [Bazzan and Grunitzki 2016]. Nessa abordagem os agentes não recebem um conjunto de rotas pré-computadas. A rota é construída ao longo da viagem. Os autores utilizaram o algoritmo de aprendizado por reforço, *Q-learning* para atualizar a sua base de conhecimento. O trabalho de [Bazzan and Grunitzki 2016] é semelhante ao de [Tumer and Agogino 2006]. Porém, neste último, os autores utilizam *difference rewards* na composição da função de recom-

pensa. Desta forma, os autores conseguem estimular a cooperação entre os agentes de modo que o congestionamento na rede seja minimizado. Estes três trabalhos têm em comum o modelo de simulação utilizado, o qual é macroscópico. O presente trabalho utiliza um modelo de simulação microscópico, que permite modelar, por exemplo entidades como semáforos e também as interações entre os agentes.

Além das técnicas que envolvem aprendizado por reforço, existem abordagens centralizadas de alocação de tráfego que alocam uma rota para cada veículo. Um método de alocação de tráfego distribui os veículos na rede levando em consideração a origem e destino dos mesmos com o objetivo de minimizar um custo associado (pode ser tempo de viagem, emissão de poluentes, entre outros). O DUA é um método de alocação de tráfego que visa calcular o *user equilibrium* [Wardrop 1952]. Especificamente, o DUA encontra uma rota para cada veículo de forma iterativa. A cada iteração, o DUA computa um conjunto de rotas mais rápidas e de rotas alternativas. Esse conjunto é coletado a partir de uma distribuição e é utilizado para decidir a rota real da próxima iteração. No entanto, o DUA difere da nossa abordagem por ser um método centralizado que depende de conhecimento completo sobre a rede. O que pode ser irrealista no processo diário de escolha de rotas dos motoristas.

4. Metodologia

4.1. Algoritmo

O algoritmo de aprendizado por reforço utilizado neste trabalho é o *Q-learning*. No início da execução do algoritmo, as k rotas entre a origem e o destino são computadas e tais rotas não mudam até o final da execução. Além disso, cada veículo mantém uma tabela Q, a qual é inicializada com zeros e atualizada ao longo dos episódios.

No primeiro episódio as rotas são escolhidas de forma aleatória, uma vez que a tabela Q só possui valores zero. Nos episódios seguintes, cada veículo escolhe uma rota de acordo com a sua tabela Q. Essa escolha é feita através da estratégia de exploração ε-gulosa (ver Seção 2). Após, cada veículo executa a sua rota escolhida.

O episódio só termina quando todos os veículos chegam no seu destino ou quando o episódio atinge um limite de execução. O último caso acontece quando a rede entra em um estado de *grid-lock*[2]. Quando o episódio termina cada veículo recebe um valor de recompensa, que é o seu tempo de viagem. Com o valor da recompensa, os veículos atualizam suas tabelas Q utilizando a Equação (1).

4.2. Cenário estudado

Como cenário de validação utilizamos uma rede em grade 3 x 3 similar à do artigo [Mannion et al. 2016], mas com algumas modificações. A capacidade da rede foi duplicada. A rede possui 18040,80 metros ao total (somatório do comprimento dos arcos) e cada veículo possui um tamanho de 5 metros. Portanto, a capacidade total da rede é de 3608 veículos.

[2]Um problema encontrado na rede de transporte é a parada total. Isso acontece quando alguns veículos não terminam de percorrer a sua rota, pois existem filas contínuas que bloqueiam todas as interseções da rede. Esse caso acontece principalmente quando a demanda da rede é muito grande, esse fenômeno é chamado de *grid-lock*.

Em relação aos pares OD, em primeiro momento considerou-se uma combinação de todos os vértices. Porém, constatamos empiricamente que manter todos os pares OD aumenta a complexidade de análise, mas não torna o problema mais complexo. Portanto, optamos por modelar apenas os pares OD mais representativos, sem que isso comprometa a generalidade da nossa abordagem, especificamente: norte-sul, leste-oeste, noroeste-sudeste e nordeste-sudoeste. A Figura 1 mostras os distritos (conjuntos de nós) selecionados como origem e/ou destino através de linhas tracejadas. No total, oito pares OD foram definidos: 0 a 8, 8 a 0, 2 a 6, 6 a 2, 1 a 7, 7 a 1, 3 a 5 e, 5 a 3.

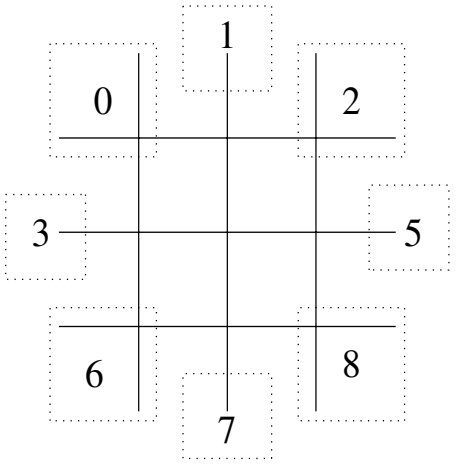

Figura 1. Rede 3 x 3 com demarcação dos distritos

Experimentos foram realizados com diferentes demandas a fim de analisar os resultados do algoritmo na rede proposta. A Tabela 1 apresenta a demanda para cada par OD e a demanda total da rede. Inicialmente, foi utilizada uma demanda com 900 veículos, que equivale a 25% da capacidade total da rede. A partir disso, outras demandas foram testadas. Vale ressaltar que, quando a demanda é muito grande, podem ocorrer eventos de *grid-lock*. Empiricamente, para a rede utilizada em nossos experimentos, constatamos que *grid-locks* ocorrem com maior frequência a partir de uma demanda de 50% da capacidade da rede. Desta forma, optamos por limitar as demandas utilizadas a até 66% (2400 veículos) da capacidade da rede.

Tabela 1. Demanda por pares OD e demanda total

Pares OD								**Total de veículos**	**Ocupação da rede %**
0-8	**8-0**	**2-6**	**6-2**	**1-7**	**7-1**	**3-5**	**5-3**		
175	175	175	175	50	50	50	50	900	25%
200	200	200	200	50	50	50	50	1000	30%
400	400	400	400	50	50	50	50	1800	50%
400	400	400	400	100	100	100	100	2000	55%
500	500	500	500	50	50	50	50	2200	61%
500	500	500	500	100	100	100	100	2400	66%

4.3. Métrica de avaliação

As métricas empregadas para avaliar o algoritmo implementado foram o tempo médio de viagem global da rede e o tempo médio de viagem por par OD. Para ambos, quanto menor for o valor, melhor é a performance do algoritmo. Os resultados gerados foram comparados entre si, nas diferentes configurações do algoritmo e com o DUA. Como o DUA é um método determinístico e, portanto, gera sempre o mesmo resultado, foi apresentado o seu resultado final após uma execução.

5. Resultados e discussão

Esta seção apresenta uma série de experimentos para analisar a eficiência do aprendizado por reforço na escolha de rotas dos veículos. Foram testadas diferentes configurações, alterando os valores da taxa de decaimento do algoritmo ε-guloso, a taxa de aprendizagem e o valor de k do algoritmo KSP. Para cada configuração proposta, o algoritmo foi executado 10 vezes e o resultado é a média dessas execuções.

O parâmetro k do algoritmo KSP foi testado com os valores 2, 3 e 4. No entanto, o que apresentou melhores resultados foi $k = 4$, pois com os outros valores os veículos só teriam 2 ou 3 rotas para escolher, sobrecarregando tais rotas. Para o algoritmo ε-guloso, o valor de ε é inicializado com 1 e, a cada episódio, é multiplicado por uma taxa de decaimento dr. Dessa maneira, há uma maior exploração das ações no início do algoritmo e com passar do tempo, há mais aproveitamento das ações. A taxa de decaimento dr foi testada com diversos valores e somente os que apresentaram resultados mais relevantes são apresentados.

5.1. Curva de aprendizagem

Um ponto importante para o algoritmo do *Q-learning* é a convergência. A Figura 2 apresenta um gráfico do tempo médio de viagem em relação aos episódios para as diferentes demandas testadas. O tempo médio de viagem está em escala logarítmica. Observa-se que há uma grande diferença de valores entre as demandas nos primeiros episódios. Quando temos uma demanda menor, 900 ou 1000 veículos, há uma competição menor pelos arcos da rede, então o algoritmo consegue encontrar uma boa distribuição para os veículos, mesmo que haja maior exploração no início. Já quando a demanda aumenta, com 1800 veículos ou mais, há uma maior competição pelos links, por isso a curva nos episódios iniciais é maior. A maior curva do gráfico é para a demanda que possui 2000 veículos que embora não seja a maior, possui uma distribuição da demanda pelos pares OD diferente das configurações com demanda maior. Porém, mesmo com valores mais altos para demandas maiores, o algoritmo se estabiliza com aproximadamente 20 episódios e converge ao longo do tempo. Para todas as demandas foram executado 150 episódios, pois o algoritmo do *Q-learning* é sensível a taxa de exploração e observou-se melhora nos resultados, mesmo que pequena, nos episódios seguintes.

5.2. Calibração dos parâmetros do algoritmo

Nesta seção descrevemos os experimentos realizados para definir a melhor configuração dos parâmetros. Dividimos a análise em dois grupos: $dr = 0,90$ (Tabela 2) e $dr = 0,99$ (Tabela 3). A diferença entre a Tabela 2 e a Tabela 3 é a taxa de decaimento. Na primeira

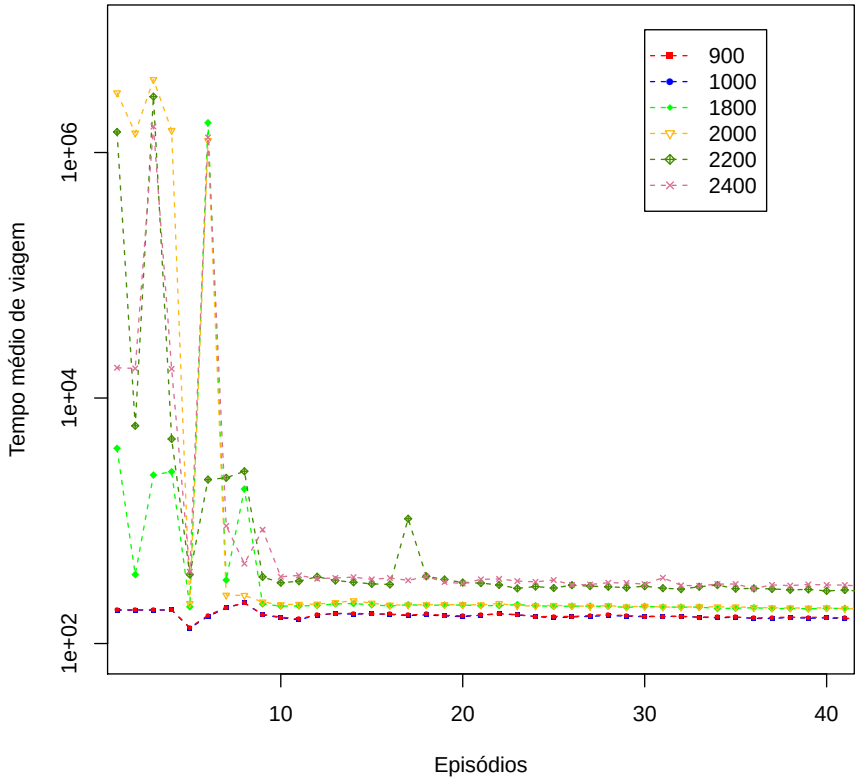

Figura 2. Curva de aprendizagem sobre o tempo médio de viagem

há um maior aproveitamento das ações, enquanto na segunda há uma maior exploração das ações. No entanto, ambas possuem resultados semelhantes.

A Tabela 2 mostra a comparação entre os diferentes valores da taxa de aprendizagem, simbolizado por α. A taxa de decaimento, simbolizada por dr, utilizada para essa tabela foi de 0,90. Os resultados apresentados são o tempo médio de viagem e o desvio padrão de cada resultado. Os melhores resultados são os destacados em negrito. Dentre os melhores resultados temos $\alpha = 0,8$ e $\alpha = 0,9$ são os que apresentam valores melhores do tempo médio de viagem. Optou-se por destacar os dois melhores resultados, porque os valores são muito próximos. Dessa maneira, aplicou-se o teste estatístico t de *Student* para verificar se os valores de $\alpha = 0,8$ e $\alpha = 0,9$ são equivalentes.

Contatou-se que com uma demanda de 900, 1000, 1800 e 2000 veículos com um nível de significância de 5%, $\alpha = 0,8$ e $\alpha = 0,9$ são equivalentes. Entretanto, para a demanda de 2200 e 2400 veículos, utilizando o mesmo intervalo de confiança, os valores diferem entre si .

Tabela 3 apresenta a comparação entre os diferentes valores da taxa de aprendizagem. A taxa de decaimento utilizada é de 0.99. Da mesma forma que a Tabela 2, os melhores valores são apresentados em negrito. Em ambas as tabelas, os melhores resultados são que possuem $\alpha = 0,8$ e $\alpha = 0,9$. Isso acontece porque quanto maior o valor da

taxa de aprendizagem, maior será a atualização com o conhecimento novo. Por exemplo, se esse valor é 1, o agente considera apenas a informação mais recente.

Tabela 2. Tempo médio de viagem (e desvio padrão) para diferentes demandas e valores de α com taxa de decaimento de 0,90

	α				
Demanda	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
900	127,7 (0,6)	122,8 (0,4)	121,2 (0,4)	**120,6 (0,5)**	**120,9 (0,3)**
1000	129 (0,9)	123,9 (0,5)	122,3 (0,8)	**121,9 (0,3)**	**122,2 (0,6)**
1800	160,9 (2,2)	154,9 (3,3)	152,9 (2,7)	**152,5 (1,6)**	**153,8 (4,1)**
2000	169 (6,2)	158,1 (4,6)	165,4 (10,7)	**160,7 (4,1)**	**158,8 (4,7)**
2200	241,4 (14,9)	234,8 (28,9)	229,9 (9,5)	**230,4 (9,5)**	**226,5 (14,0)**
2400	238,3 (19,4)	222,6 (19,9)	228,4 (17,9)	**220,3 (19,0)**	**216,3 (15,2)**

Tabela 3. Tempo médio de viagem (e desvio padrão) para diferentes demandas e valores de α com taxa de decaimento de 0,99

	α				
Demanda	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**
900	127,4 (0,5)	122,7 (0,5)	121,2 (0,4)	**120,9 (0,3)**	**120,7 (0,5)**
1000	129,3 (0,8)	124,1 (1,0)	122,8 (0,7)	**122,1 (0,3)**	**122,4 (0,5)**
1800	161,3 (3,6)	154,8 (3,1)	**153,9 (2,1)**	**153,5 (2,2)**	154,4 (4,5)
2000	165,9 (7,5)	163,8 (8,6)	160,4 (8,1)	**158,8 (3,5)**	**158,4 (5,2)**
2200	214,4 (23,1)	224,7 (15,7)	245,2(28,7)	**242,5 (26,3)**	**228,5 (17,0)**
2400	222,6 (11,0)	230,7(23,1)	227,4 (32,0)	**223,5 (17,7)**	**222,7 (14,1)**

5.3. Comparação com o DUA

Nesta seção, apresentamos uma comparação dos resultados obtidos pelo nosso método com aqueles obtidos pelo método DUA. Para tanto, utilizou-se a configuração dos melhores resultados anteriores. Para a taxa de decaimento, optou-se por utilizar tanto 0,90 quanto 0,99, dado que ambas apresentaram resultados estatisticamente equivalentes. Para o α, os melhores resultados foram com 0,8 e 0,9. No entanto, optou-se $\alpha = 0,8$, que propicia um melhor aproveitamento do conhecimento novo sobre o conhecimento antigo.

A Tabela 4 mostra a comparação das duas configurações (taxa de decaimento 0,90 e 0,99). Para ambas, foi utilizado $\alpha = 0.8$ e $k = 4$. Os resultados do DUA para as demandas de 2200 e 2400 veículos não são mostrados porque o mesmo não foi capaz de concluir sua execução após o tempo limite de duas horas.

Conforme a Tabela 4, para uma demanda de 900 e 1000 veículos, tanto para $dr = 0,90$ quanto para $dr = 0,99$, os resultados obtidos pelo nosso método são aproximadamente 15% melhores que aqueles obtidos pelo DUA. Já para a demanda de 1800 veículos $dr = 0,90$ e $dr = 0,99$ são aproximadamente 13% melhores que o DUA. Porém, quando temos uma demanda maior, com 2000 veículos, os experimentos com $dr = 0,90$ e $dr = 0,99$ são 4% e 5% melhores que o DUA, respectivamente. Portanto, verificou-se que, com $dr = 0,90$, nosso método é em média 11,9% melhor que o DUA e com o $dr = 0,99$, nosso método é em média 11,7% melhor que o DUA.

Tabela 4. Tempo médio de viagem (e desvio padrão) dos resultados obtidos pelo nosso método e pelo DUA, para diferentes demandas e taxas de decaimento

Demanda	DUA	$dr = 0,90$	$dr = 0,99$
900	141,9	120,6 (0,5)	120,9 (0,3)
1000	143,6	121,9 (0,3)	122,1 (0,3)
1800	176,2	152,5 (1,6)	153,5 (2,2)
2000	167,3	160,7 (4,1)	158,8 (3,5)
2200	-	230,4 (9,5)	242,5 (26,3)
2400	-	220,3 (19,0)	223,5 (17,7)

5.4. Comparação entre pares OD

Mesmo que os resultados do tempo médio de viagem global da nossa abordagem tenham sido melhores que os do DUA, não quer dizer que em todos os aspectos o tempo de viagem diminui. Isto ocorre porque muitas vezes pode ser dado maior vazão para uma via, enquanto outra esta completamente parada. Dessa forma, faz-se necessário observar o tempo médio de viagem por par OD.

As Tabelas 5 e 6 apresentam os resultados obtidos pelo DUA e pelo nosso método por par OD. Nesse caso, podemos observar que nos pares 0-8, 8-0, 2-6 e 6-2, os resultados do nosso algoritmo possuem valores melhores se comparados ao DUA. No entanto, para os pares OD 1-7, 7-1, 3-5 e 5-3, o tempo médio de viagem resultante do DUA é melhor que na nossa abordagem. Isto ocorre porque o DUA encontra um equilíbrio local, melhorando a rota para alguns pares OD, mas não em termos gerais.

Tabela 5. Tempo médio de viagem por par OD para as demandas de 900 e 1000 veículos

Pares OD	900			1000		
	DUA	$\boldsymbol{dr = 0,90}$	$\boldsymbol{dr = 0,99}$	**DUA**	$\boldsymbol{dr = 0,90}$	$\boldsymbol{dr = 0,99}$
0-8	161,5	127,7 (0,7)	127,6 (0,5)	159,9	130 (0,9)	130 (1,4)
8-0	150,2	127,5 (0,8)	127,4 (0,8)	153,1	125,8 (0,6)	125,9 (1,1)
2-6	150,5	124,9 (0,6)	125,5 (0,5)	159,3	125,4 (0,5)	125,3 (0,5)
6-2	160,7	127,8 (0,6)	128,2 (0,8)	157,8	131,2 (0,4)	131,1 (0,6)
1-7	92,14	101 (0,7)	100,7 (0,7)	95,9	98 (0,0)	98,3 (0,9)
7-1	108,4	103,1 (2,0)	103,2 (1,5)	100,6	102,3 (1,3)	103 (1,5)
3-5	91,6	97,1 (0,3)	97,2 (0,6)	90,1	92,2 (0,6)	92,3 (0,7)
5-3	90,1	93,7 (0,8)	94,2 (0,9)	89,4	97,4 (1,2)	97,6 (0,8)

6. Conclusão e trabalhos futuros

O presente trabalho apresentou uma abordagem baseada em agentes para resolver o problema da escolha de rotas. O principal objetivo foi investigar uma forma de diminuir o tempo de viagem dos veículos e melhorar o fluxo da rede. Foi utilizada uma abordagem baseada em aprendizado por reforço na escolha de rotas. No algoritmo proposto, as rotas são pré-computadas e não mudam até o fim da execução. Os agentes utilizam apenas o conhecimento adquirido através da sua interação com o ambiente para realizar a suas escolhas. Para validar nossa abordagem, utilizamos uma simulação microscópica de uma

Tabela 6. Tempo médio de viagem por par OD para as demandas de 1800 e 2000 veículos

Pares OD	1800			2000		
	DUA	$dr = 0,90$	$dr = 0,99$	**DUA**	$dr = 0,90$	$dr = 0,99$
0-8	176,3	164,4 (1,3)	165,9 (4,5)	183,7	185,9 (10)	184,1 (11,4)
8-0	194,0	153,4 (2,5)	154,3 (2,5)	188,1	162,3 (4,2)	160,5 (3,8)
2-6	202,1	155,1 (3,2)	154,9 (2,4)	192,4	169 (6,4)	165,9 (5,7)
6-2	188,2	159,6 (3,1)	161,8 (4,0)	185,2	168,9 (4,2)	166,7 (3,4)
1-7	96,6	103,4 (2,8)	102,3 (2,2)	103,9	108,8 (5,8)	107,1 (3,8)
7-1	103,6	112,1 (7,8)	112,8 (4,3)	106,9	123,6 (9,0)	122,3 (5,1)
3-5	96,9	105,6 (9,3)	102,1 (2,3)	97,0	106,2 (5,6)	106,4 (3,9)
5-3	96,9	109,9 (5,5)	109,9 (4,3)	97,6	130,5 (5,3)	129,9 (6,4)

rede em grade 3 x 3. Os resultados gerados foram comparados entre si, com diferentes configurações do algoritmo e com o DUA.

A presente abordagem apresentou melhores resultados que o DUA no ponto de vista global da rede. Por outro lado, quando analisamos os resultados por par OD, em alguns pares OD o DUA possui melhores resultados que a nossa abordagem. No entanto, o nosso método é melhor nos pares OD que possuem mais arcos e maior demanda. Com base nos resultados, pode-se afirmar que o aprendizado por reforço na escolha de rotas é uma boa alternativa para os problemas de tráfego enfrentados hoje.

Como trabalhos futuros, uma proposta é adicionar outros tipos de agentes na rede de transporte. Por exemplo, as interseções, que atualmente possuem semáforos com ciclo fixo, podem ser tratadas como agentes. Dessa maneira, pode-se aplicar aprendizado por reforço nos semáforos, permitindo que estes agentes controlem as interseções de forma mais eficiente e adaptativa, melhorando o fluxo nos cruzamentos.

Referências

Bazzan, A. L. C. and Grunitzki, R. (2016). A multiagent reinforcement learning approach to en-route trip building. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 5288–5295.

Bazzan, A. L. C. and Klügl, F. (2013). *Introduction to Intelligent Systems in Traffic and Transportation*, volume 7 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan and Claypool.

Behrisch, M., Bieker, L., Erdmann, J., and Krajzewicz, D. (2011). SUMO - simulation of urban mobility: An overview. In *SIMUL 2011, The Third International Conference on Advances in System Simulation*, pages 63–68, Barcelona, Spain.

Dia, H. and Panwai, S. (2014). *Intelligent Transport Systems: Neural Agent (Neugent) Models of Driver Behaviour*. LAP Lambert Academic Publishing.

Dias, J. C., Machado, P., Silva, D. C., and Abreu, P. H. (2014). An inverted ant colony optimization approach to traffic. *Engineering Applications of Artificial Intelligence*, 36(0):122–133.

Galib, S. M. and Moser, I. (2011). Road traffic optimisation using an evolutionary game.

In *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*, GECCO '11, pages 519–526, New York, NY, USA. ACM.

Mannion, P., Duggan, J., and Howley, E. (2016). An experimental review of reinforcement learning algorithms for adaptive traffic signal control. In McCluskey, L. T., Kotsialos, A., Müller, P. J., Klügl, F., Rana, O., and Schumann, R., editors, *Autonomic Road Transport Support Systems*, pages 47–66. Springer.

Narendra, K. S. and Thathachar, M. A. L. (1989). *Learning Automata: An Introduction*. Prentice-Hall, Upper Saddle River, NJ, USA.

Ortúzar, J. d. D. and Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons, Chichester, UK, 4 edition.

Ramos, G. de. O. and Grunitzki, R. (2015). An improved learning automata approach for the route choice problem. In Koch, F., Meneguzzi, F., and Lakkaraju, K., editors, *Agent Technology for Intelligent Mobile Services and Smart Societies*, volume 498 of *Communications in Computer and Information Science*, pages 56–67. Springer Berlin Heidelberg.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Tumer, K. and Agogino, A. (2006). Agent reward shaping for alleviating traffic congestion. In *Workshop on Agents in Traffic and Transportation*, Hakodate, Japan.

Wardrop, J. G. (1952). Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II*, 1(36):325–362.

Yen, J. Y. (1971). Finding the k shortest loopless paths in a network. *Management Science*, 17(11):712–716.