Timóteo Fonseca Santos

Q-*learning* baseado em pedágios com pagamento circunstancial

Manaus, AM - Brasil

2023

Timóteo Fonseca Santos

Q-*learning* baseado em pedágios com pagamento circunstancial

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas como parte dos requisitos necessários para a obtenção do grau de Mestre em Informática.

Universidade Federal do Amazonas – UFAM

Instituto de Computação – ICOMP

Programa de Pós-Graduação em Informática – PPGI

Orientador: Prof. Dr. Moisés Gomes de Carvalho

Manaus, AM - Brasil

2023

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Santos, Timóteo Fonseca
S237q Q-learning baseado em pedágios com pagamento circunstancial / Timóteo Fonseca Santos . 2023
81 f.: il. color; 31 cm.

Orientador: Moisés Gomes de Carvalho
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Aprendizagem de máquina. 2. MCT - Marginal-cost tolling. 3. Q-learning. 4. TQ-learning. 5. Congestionamento de tráfego. I. Carvalho, Moisés Gomes de. II. Universidade Federal do Amazonas III. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"Q-LEARNING BASEADO EM PEDÁGIOS COM PAGAMENTO CIRCUNSTANCIAL"

TIMÓTEO FONSECA SANTOS

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Eduardo James Pereira Souto - PRESIDENTE

Prof. Dr. David Braga Fernandes de Oliveira - MEMBRO INTERNO

Prof. Dr. Mário Salvatierra Júnior - MEMBRO EXTERNO

Manaus, 25 de julho de 2023

Documento assinado eletronicamente por **David Braga Fernandes de Oliveira**, **Professor do Magistério Superior**, em 09/08/2023, às 14:45, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Eduardo James Pereira Souto**, **Professor do Magistério Superior**, em 09/08/2023, às 16:44, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Mário Salvatierra Júnior**, **Professor do Magistério Superior**, em 11/08/2023, às 15:52, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1592005** e o código CRC **583BD23C**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.030908/2023-70 SEI nº 1592005

Dedico este trabalho a Santo Tomás de Aquino, padroeiro dos acadêmicos, sem cuja intercessão a conclusão desta obra não teria sido possível. Aos meus pais e familiares, que padeceram junto comigo os desafios que foram superados durante o percurso desse longo caminho. A todos os amigos e colegas que de alguma forma me ajudaram.

Agradecimentos

A Deus, que me forneceu todos os meios de que precisei para a conclusão deste trabalho, abrindo portas quando os desafios pareciam insuperáveis e iluminando minha mente todas as vezes em que eu não via solução. A imensidão de Sua generosidade não cabe em palavras, e sem Ele nada teria sido possível. Também aos dois grandes intercessores ao Seu lado a quem nunca deixei de recorrer: Santo Tomás de Aquino, padroeiro dos acadêmicos e meu modelo de vida intelectual, e Santo José de Cupertino, padroeiro dos estudantes.

Aos membros da minha família que oraram continuamente por mim, me apoiaram, me sustentaram e me acompanharam ao ponto de compartilharmos juntos a fadiga do processo. Em especial à minha mãe, Daniela Rodrigues Fonseca Lyra, que esteve sempre comigo e inúmeras vezes me motivou a seguir em frente.

Ao meu grande amigo, Rúben Jozafá Silva Belém, meu companheiro de lutas desde o ensino médio, que esteve à minha frente em nossos respectivos mestrados de informática e assim pôde me mostrar o caminho das pedras. Pelos compartilhamento de momentos de desafios e encorajamento mútuo. Desde o começo as suas ajudas, grandes ou pequenas, foram algumas das mais cruciais que obtive ao longo da jornada.

Ao meu psicólogo, Antônio da Conceição Marinho, cujas sessões muito me ajudaram a me recobrar quando minhas energias mais pareciam esgotadas. Com seus conselhos e os diversos planos de ação que traçamos juntos, pude manter minha motivação e equilibrar a produção deste trabalho com as demais metas do meu dia-a-dia. Sua ajuda profissional permitiu que eu obtivesse mais auto-conhecimento a cada desafio que eu superava neste trabalho.

Ao meu orientador, Prof. Dr. Moisés Gomes de Carvalho, pela oportunidade de realizarmos este trabalho juntos, pelo encorajamento, pelo estabelecimento de metas e limites, pelas nossas reuniões em que tudo se tornava mais claro e os próximos passos ficavam muito bem definidos. Sinto-me honrado por essa orientação.

Ao Programa de Pós-Graduação de Informática e a todos os professores que participaram da minha jornada, por toda ajuda e compreensão que me ofereceram.

A Gabriel Ramos de Oliveira, autor de artigos que foram basilares para este trabalho, pela sua disponibilidade em se comunicar conosco para esclarecer dúvidas, compartilhar dicas e a imensa solicitude em nos fornecer acesso a repositórios que serviram como ponto de partida para a implementação do modelo que apresentamos.

"Começar é de todos; perseverar, de santos.
Que tua perseverança não seja consequência cega do primeiro impulso, fruto da inércia;
que seja uma perseverança refletida."
(Santo Josemaria Escrivá)

Resumo

Congestionamentos são um problema recorrente nas grandes cidades, resultando em perda de produtividade, poluição e diminuição da qualidade de vida. As técnicas existentes para resolução de congestionamentos de tráfego nem sempre são eficazes ou economicamente viáveis. No entanto, a implementação de sistemas de pedágio para controlar o fluxo de tráfego em áreas movimentadas já chegou a mostrar melhorias observáveis. A análise matemática e a simulação virtual surgem como ferramentas úteis para avaliar o custo-benefício de cada abordagem. Mitigar congestionamentos envolve equilibrar o desempenho ideal do sistema e o equilíbrio do usuário, exigindo incentivos para tornar alinhados o comportamento individual do motorista com melhorias no desempenho do sistema. Há uma ampla base teórica apoiando a eficácia de abordagens baseadas em pedágio. No entanto, a premissa comum de que todos os motoristas pagam pedágio pode limitar a eficiência dos modelos no mundo real devido a evasão ou limitações econômicas. Abordando esses desafios, este trabalho explora os impactos de diferentes níveis e modos de participação em sistemas de pedágio. Adaptamos uma abordagem baseada em pedágio existente para introduzir diversos cenários de pagamento seletivo de pedágio, e investigamos a viabilidade da adoção gradual de sistemas tarifários. Tendo implementado uma variação do algoritmo TQ-*learning*, conseguimos controlar parâmetros como a proporção de motoristas que pagam pedágio ou a proporção de ruas mais movimentadas onde o pedágio é obrigatório. Por meio de experimentos em múltiplas proporções, apresentamos resultados que ampliam a base de conhecimento para a tomada de decisão prática na resolução de congestionamentos. Nossas descobertas demonstram que quando o sistema de pedágio é implementado gradualmente por meio do crescente número de usuários que pagam regularmente, os ganhos são constantes e sem introduzir comportamento caótico. No entanto, ao introduzir pedágios por elos ou rotas dos mais aos menos movimentados, os resultados foram, na melhor das hipóteses, inconclusivos e, na pior, provocaram uma deterioração do desempenho do sistema em comparação com a ausência de implementação.

Palavras-chaves: Aprendizagem de máquina, MCT, Q-learning, TQ-learning, congestionamento de tráfego

Abstract

Congestion is a recurring problem in large cities, leading to productivity loss, pollution, and decreased quality of life. Existing techniques for traffic congestion resolution are not always effective or economically viable. However, implementing toll systems to control traffic flow in busy areas has shown observable improvements. Mathematical analysis and virtual simulation emerge as useful tools to evaluate the cost-effectiveness of each approach. Mitigating congestion involves balancing the optimal system performance and user equilibrium, requiring incentives to align individual driver behavior with system improvements. Toll-based approaches have a theoretical foundation in effectively addressing this issue. However, the assumption that all drivers pay tolls may limit the real-world efficiency of the models due to non-compliance or economic limitations. Addressing these challenges, this work explores the impacts of different levels and modes of participation in toll systems. We adapt an existing toll-based approach to handle diverse scenarios of selective toll payment and investigate the viability of the gradual adoption of toll systems. Having implemented a variation of the TQ-learning algorithm with conditional payment, we can control parameters such as the proportion of toll-paying drivers or the proportion of roads where tolling is obligatory. Through experiments at multiple proportions, we present results that expand the knowledge base for practical decision-making in congestion resolution. Our findings demonstrate that when the toll system is gradually implemented through an increasing proportion of regularly-paying users the gains are steady without introducing chaotic behavior. However, when introducing tolling on a per-route or per-link basis the results were at best inconclusive and, at worst, they caused a deterioration of system performance compared to no implementation.

Key-words: Machine learning, MCT, Q-learning, TQ-learning, traffic congestion

Lista de ilustrações

Lista de tabelas

Lista de Algoritmos

Lista de abreviaturas e siglas

ACO	*Ant Colony Optimization* ("Otimização por Colônia de Formigas")
EM	Eastern-Massachusetts
IACO	*Inverted Ant Colony Optimization* ("Otimização por Colônia de Formigas Invertida")
MCT	*Marginal-Cost Tolling* ("Tarifagem por Custo Marginal")
OD	Origem-Destino
PoA	*Price of Anarchy* ("Preço da Anarquia")
Q-*learning*	*Q(uality) learning*
SO	*System Optimum* ("Ótimo do Sistema")
TQ-*learning*	*Toll-based Q-learning*
UE	*User Equilibrium* ("Equilíbrio de Usuários")

Lista de símbolos

a	Uma ação
$a_{i,t}$	Ação do motorista i no episódio t
A	Conjunto de todos os conjuntos de ações
A_i	Conjunto de ações de um motorista i
$\mathbb{B}$	Conjunto dos booleanos
c	Equações de custo básicas
c_l	Equação de custo do elo l
C	Equações de custo
C_R	Equação de custo da rota R
$C_{i,R}$	Equação de custo da rota R para o motorista i
d	Quantidade total de motoristas
D	Conjunto de motoristas
f	Equações de fluxo
f_l	Equação de fluxo do elo l
G	Grafo da malha de tráfego
i	Motorista i
K	Quantidade de rotas mais curtas a serem calculadas pra cada par OD
l	Elo l
L	Conjunto de elos
m	Modo de pagamento
M	Conjunto de valores de permutação dos parâmetros υ e ρ
N	Conjunto de nós
$\mathbb{N}$	Conjunto dos números naturais

$\mathbb{N}_0$	Conjunto dos números naturais incluindo 0
p_n	Subcondição n de pagamento de pedágio
P	Problema de tráfego
$\mathcal{P}$	Condição de pagamento de pedágio
Q	Tabela-Q
Q_t	Tabela-Q no momento t
$Q_{i,t}$	Tabela-Q do motorista i no episódio t
r	Equação de recompensa
R	Rota R
$\mathbb{R}$	Conjunto dos números reais
$\mathbb{R}^+$	Conjunto dos números reais positivos
t	Episódio t
T	Quantidade total de episódios
$\mathcal{U}_X$	Valor aleatório tirado de uma distribuição uniforme X
w	Tempos médios de viagem para cada episódio de um experimento
w_t	Tempo médio de viagem no episódio t
w_T	Tempo médio de viagem no último episódio
x	Conjunto de quantidades de motoristas (ou volumes de tráfego) por elo
x_l	Quantidade de motoristas (ou volume de tráfego) no elo l
$\mathbb{Z}$	Conjunto dos números inteiros
v	Tempo médio de viagem final
$\overline{v}$	Média de valores v obtidos em repetidos experimentos para uma dada configuração experimental
$\overline{v}_{\mathrm{UE}}$	Valor $\overline{v}$ no caso de aproximação de UE
$\overline{v}_{\mathrm{SO}}$	Valor $\overline{v}$ no caso de aproximação de SO
$\Delta\overline{v}$	Diferença entre valores $\overline{v}$ de casos sucessivos do mesmo parâmetro (υ ou ρ) em uma mesma configuração experimental

α	Fator de aprendizado
γ	Malha de tráfego
ϵ	Fator de exploração
λ	Taxa de decaimento de α
λ^t	Valor de α no episódio t
μ	Taxa de decaimento de ϵ
μ^t	Valor de ϵ no episódio t
ξ	Conjunto com todos os elos de L ordenados do maior ao menor valor x_l
ϖ	Booleano que indica se um pedágio deve ser pago em c_l
ρ	Parâmetro de controle da proporção de elos mais movimentados
ϱ	Quantidade de elos acima do limiar estabelecido por ρ
σ	Quantidade de repetições de experimentos para cada configuração experimental
ς	Conjunto com os ϱ elos mais movimentados de L
τ	Equações de pedágio
τ_l	Equação de pedágio do elo l
υ	Parâmetro de controle da proporção de motoristas que são usuários
ϕ	Fórmula de proximidade

Sumário

1 Introdução

Congestionamentos são um problema recorrente em grandes cidades, resultando em perda de produtividade (Somuyiwa et al., 2015), poluição e diminuição na qualidade de vida (Zhong et al., 2017). As técnicas existentes para resolver congestionamentos no trânsito nem sempre são eficazes ou economicamente viáveis. Por exemplo, ampliar capacidade de ruas pode piorar o fluxo de tráfego através do efeito de demanda induzida (Wood et al., 1994). No entanto, a implementação de sistemas de pedágio para controlar o fluxo entrando e saindo de áreas movimentadas tem demonstrado melhorias observáveis, como evidenciado em Londres (Leape, 2006). Dados os desafios e custos associados à investigação das técnicas de alívio de congestionamento, a análise matemática e a simulação virtual surgem como ferramentas úteis para avaliar o custo-benefício de cada abordagem.

Congestionamentos surgem quando o volume de tráfego em uma via causa uma demanda por espaço maior do que ela tem disponível. Há casos em que congestionamentos podem ser mitigados através da redistribuição de motoristas ao longo de rotas diferentes, e a tarefa de encontrar tais soluções constitui um "problema de alocação de tráfego", ou *traffic assignment problem*[1]. Isso é possível quando o sistema tem um desempenho ótimo melhor do que seu desempenho no equilíbrio de usuários, isto é, quando todos os motoristas estão escolhendo as rotas que já trazem o melhor benefício individual.

Nesses casos, porém, os motoristas dificilmente farão tal redistribuição espontaneamente porque, caso já estejam realizando as melhores escolhas possíveis, mudar de estratégia trará prejuízos para si mesmos, ainda que isso beneficie o sistema como um todo. Supondo que motoristas não estão dispostos a se sacrificar pelos demais, são necessários incentivos que equilibrem esses possíveis prejuízos. Abordagens com pedágios têm amplo embasamento teórico como solução eficaz nesse contexto (Pigou, 1920; Hearn and Ramana, 1998). Elas são capazes de mitigar congestionamentos cobrando dos motoristas pedágios proporcionais ao quanto de prejuízo causam aos demais, e motoristas buscando minimizar o valor gasto em pedágios tendem assim ao comportamento que leva ao ótimo do sistema.

Essas abordagens geralmente pressupõem que todos os motoristas do sistema estarão pagando pedágios. Na prática, entretanto, vários fatores podem limitar a cobertura do sistema de pedágios, o que prejudicaria a eficiência das soluções. Por exemplo, é improvável que uma adesão plena seja alcançada imediatamente. O sistema precisaria ser implantado e expandido gradualmente, em áreas específicas. Além disso, uma taxa de evasão entre os motoristas pode sempre estar presente. Adicionalmente, limitações econô-

[1] Também conhecido na literatura como problema de alocação de rotas (*route assignment problem*) ou problema de escolha de rotas (*route choice problem*).

micas podem impossibilitar o crescimento do sistema além de certo ponto. Os custos de implantação podem resultar em retornos cada vez menores, fenômeno conhecido como lei de rendimentos decrescentes[2]. Essas são apenas algumas das questões que podem surgir.

Outro aspecto a ser considerado é que muitas abordagens de pedágio buscam ser descentralizadas, visando ser mais adaptáveis, robustas e viáveis para implantação do que as abordagens centralizadas. No entanto, se a garantia de participação depender exclusivamente de meios centralizados, todas essas vantagens podem ser facilmente neutralizadas. É importante para uma abordagem com enfoque descentralizado desenvolver meios sustentáveis para garantir a participação do maior número de usuários. Por exemplo, identificando aspectos do modelo que favorecem a adesão voluntária ao sistema.

A investigação dos efeitos da adesão parcial em um sistema de pedágio pode ampliar nosso conhecimento deste, nos preparando para lidar com os problemas supracitados. Por exemplo, podemos determinar seu nível de tolerância à evasão e analisar se a implementação gradual resulta em melhorias consistentes de desempenho. Além disso, devemos avaliar se a implementação gradual apresenta riscos de instabilidade no sistema ou uma queda significativa no desempenho.

Neste trabalho exploramos os impactos que diferentes níveis e modos de participação podem ter em um sistema de pedágios. Criamos uma variação do algoritmo TQ-*learning* (Ramos et al., 2020a) com pagamento de pedágios circunstancial. Por "pedágios circunstanciais" ou "pagamento circunstancial" queremos dizer que, a depender das circunstâncias, o pedágio é pago ou não, o que é controlado por uma condição de pagamento.

A condição de pagamento é flexível o suficiente para nos permitir testar três cenários com circunstâncias diferentes de pagamento. Com nosso algoritmo realizamos experimentos permutando diversos casos e apresentamos como resultados os desempenhos do sistema. Dessa forma, esperamos expandir o acervo de informações relevantes para tomadas de decisões na resolução prática de congestionamentos.

O "pedágio" de que falamos não necessariamente modela um sistema com cancelas bloqueando a entrada de vias mediante cobrança, que é a situação mais familiar para brasileiros, apesar de nossa condição de pagamento permitir a simulação de situações parecidas (ver Seção 5.4). Ainda assim, a finalidade principal do pedágio não é a manutenção de vias, mas minimizar engarrafamentos. Ele também permite modelar pedágios que são cobrados automaticamente por dispositivos instalados no veículo capazes de rastrear o percurso do motorista, o que foi a sugestão original de Ramos et al. (2020a). O importante é poder modelar diversas formas de se cobrar pedágios, não se limitando ao que é aqui mencionado.

Buscamos tratar as seguintes questões de pesquisa: é possível adaptar uma abor-

[2] Conhecido em inglês como *law of diminishing returns*.

dagem baseada em pedágios para lidar com múltiplos cenários de participação parcial? Sendo possível, os resultados apoiam a adesão gradual ao sistema de pedágios? Para ambas as questões a resposta é positiva, mas a segunda tem ressalvas: apenas um dos cenários, o controle de motoristas usuários vs. não-usuários, fornece uma base realista e com bons resultados apoiando a adesão gradual de pedágios. Os demais obtiveram resultados na melhor das hipóteses inconclusivos; na pior, são evidências de que seus métodos de controle podem trazer prejuízos.

A principais contribuições deste trabalho são:

1. Uma ferramenta que, diferente das demais encontradas na literatura, permite a investigação dos efeitos de pedágios aplicados apenas parcialmente. Três formas de controle foram implementadas:
   a) Pela proporção de motoristas que são usuários e, portanto, pagam pedágios; sendo os demais não-usuários que não pagam pedágios.
   b) Pelos elos mais movimentados do sistema dentro de um limiar, sendo que o motorista que passar por pelo menos um elo dentro do limiar paga um pedágio valendo pelo o trajeto inteiro.
   c) Pelos elos mais movimentados do sistema dentro de um limiar, mas dessa vez o motorista paga um pedágio calculado apenas para os elos dentro do limiar pelos quais ele passou.
2. Uma análise de custo-benefício mais aprofundada para o TQ-*learning*.
3. Evidências de que a aplicação apenas parcial do TQ-*learning* não apresenta grandes riscos se for controlada por usuários e não-usuários.
4. Evidências de que a aplicação apenas parcial do TQ-*learning* controlando pela travessia por elos mais movimentados apresenta riscos ao desempenho do sistema.

O restante desta dissertação está organizado da seguinte forma: no Capítulo 2 definimos conceitos básicos para a compreensão do nosso trabalho e dos demais. No Capítulo 3 apresentamos trabalhos e algoritmos relacionados ao problema da solução de congestionamentos em sistemas de tráfego. No Capítulo 4 apresentamos nosso método proposto para investigar como diferentes taxas de adesão influenciam na solução de congestionamentos pelo método de tarifagem. No Capítulo 5 analisamos os resultados obtidos pela experimentação do método. Por fim, no Capítulo 6 encerramos com nossas conclusões sobre o trabalho e sugerimos direções para trabalhos futuros.

2 Referencial teórico

Nesta seção apresentamos os conceitos mais importantes para a compreensão do modelo apresentado em nosso trabalho e também dos trabalhos relacionados. Começamos definindo algumas convenções notacionais na Seção 2.1. Falamos sobre teoria de jogos e seus conceitos relevantes na Seção 2.2. Na Seção 2.3 apresentamos o problema de alocação de tráfego e explicamos conceitos de sistemas de tráfego. Na Seção 2.4 introduzimos a classe de algoritmos de aprendizagem por reforço, incluindo o Q-*learning*. Na Seção 2.5, a mais extensa, apresentamos o TQ-*learning*, a variante de Q-*learning* que é nosso *baseline*. Por fim, na Seção 2.6 encontra-se a fórmula pro cálculo de proximidade, que será importante para nossos experimentos.

2.1 Convenções notacionais

Nesta seção definimos e esclarecemos algumas das convenções notacionais que serão usadas ao longo do trabalho, visando eliminar ambiguidade.

2.1.1 Intervalos

Intervalos numéricos fechados serão denotados por colchetes representando intervalos fechados (e.g. $[a, b]$). Parênteses de um lado ou outro representarão pontos em que o intervalo é aberto (e.g. $[a, b)$ ou $(a, b]$). Para evitar confusão com tuplas, intervalos abertos tanto à esquerda quanto à direita serão representados não por parênteses dos dois lados – i.e. (a, b) – mas por colchetes invertidos; ou seja: $]a, b[$. Em intervalos contínuos os pontos serão separados por vírgula e em intervalos discretos serão separados por dois pontos finais (i.e. $[a..b]$). Exemplos:

1. $[a, b] \Rightarrow \{x \in \mathbb{R} : a \leqslant x \leqslant b\}$
2. $]a, b[\Rightarrow \{x \in \mathbb{R} : a < x < b\}$
3. $[a..b) \Rightarrow \{x \in \mathbb{Z} : a \leqslant x < b\}$

2.1.2 Valores aleatórios

$\mathcal{U}_X$ denotará um valor aleatório tirado de uma distribuição uniforme X, seja esta contínua ou discreta. X pode ser um intervalo numérico ou um conjunto. Exemplos:

1. $\mathcal{U}_{[a,b)}$ representa um valor aleatório $x \in \mathbb{R}$ onde $a \leqslant x < b$;

2. Seja $S = \{a, b, c\}$, $\mathcal{U}_S$ representa um valor aleatório $x \in S$ que pode ser qualquer elemento de S com igual probabilidade entre eles.

2.1.3 Booleanos e colchetes de Iverson

O conjunto de valores booleanos, que podem ser apenas verdadeiros ou falsos, será expresso pela notação $\mathbb{B}$.

Usaremos a notação conhecida como colchetes de Iverson para converter proposições lógicas em valores numéricos, i.e. $\mathbb{B} \to \{0, 1\}$. Mais especificamente: caso um valor booleano P seja verdadeiro, ele se torna 1; caso seja falso ($\neg P$), ele se torna 0. Ela é definida na Equação 2.1 seguindo a especificação de Knuth (1992).

$$[P] \Rightarrow \begin{cases} 1, & \text{se } P \text{ for verdadeiro;} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.1}$$

2.2 Teoria de jogos

Alguns dos conceitos adotados neste trabalho têm origem na teoria de jogos, que é o estudo de modelos matemáticos de conflito e interação entre tomadores de decisão racionais (Myerson, 1997). Os motoristas da simulação buscam minimizar os seus custos de viagem (por exemplo, tempo e tarifas) como agentes racionais e auto-interessados, ou seja, seu objetivo é maximizar os benefícios próprios mesmo que isso prejudique os demais, gerando competição entre si. Portanto, seu comportamento pode ser analisado pela perspectiva da teoria de jogos.

O resultado esperado é que os agentes cheguem a um equilíbrio no sistema onde nenhum deles obtém qualquer vantagem por escolher uma estratégia diferente da consolidada – ou seja, o melhor desempenho a nível de indivíduos. Esse ponto é chamado de "equilíbrio de usuários" (*User Equilibrium* – UE) (Wardrop, 1952), e é equivalente ao "equilíbrio de Nash" (Nash, 1951) da teoria dos jogos. No nosso modelo o UE é quantificado como um valor real positivo correspondente ao tempo médio de viagem final obtido por motoristas, que será referido daqui pra frente como "valor de UE". Quanto maior o tempo médio de viagem, pior o desempenho do sistema.

O melhor desempenho a nível de sistema, ou "ótimo de sistema" (*System Optimum* – SO), será compreendido no nosso modelo como o mínimo possível que o tempo médio de viagem no sistema pode atingir. O SO é quantificado da mesma forma que o UE, tendo um "valor de SO" também. Eles nem sempre são equivalentes, podendo o valor de UE ser igual ou pior do que o valor de SO (UE $\geqslant$ SO), como demonstrado no exemplo do paradoxo de Braess (1968). Nos casos em que o valor de UE é pior, o SO só pode

ser atingido se os agentes estiverem dispostos a tomarem escolhas sub-ótimas sob uma perspectiva individual, o que é incompatível com seu comportamento auto-interessado.

Apesar de existirem valores de UE e SO ideais para uma rede, veremos adiante diversos trabalhos (além do nosso) que realizam aproximações desses valores. Chamaremos esses valores aproximados de "aproximação de UE/SO", a depender do caso que está sendo aproximado.

Por fim, é conhecida como "preço da anarquia" (*Price of Anarchy* – PoA) a deterioração da eficiência do sistema por conta do comportamento "egoísta" de usuários (Koutsoupias and Papadimitriou, 1999). O PoA é a razão entre os valores de UE e de SO, ou seja, $\text{PoA} = \frac{\text{UE}}{\text{SO}}$, onde $\text{PoA} \geqslant 1$. Consideramos que é mais eficiente o sistema quanto mais próximo de 1 for seu PoA.

2.3 Sistemas de tráfego

Um sistema de tráfego pode ser compreendido como um grafo que visa modelar rotas, com elos representando as ruas, vértices sendo intercessões e destinos, e rotas formadas por elos interligando os vértices.

O problema essencial do nosso trabalho é alocar tráfego a fim de minimizar o congestionamento no sistema. A minimização de congestionamentos em um sistema é equivalente a convergir o desempenho dele ao seu SO, fazendo com que o PoA seja 1, sendo seu desempenho medido pelo tempo médio de viagem dos motoristas, o que é calculado a partir de funções de fluxo, que são apresentadas na Subseção 2.3.1. Um conceito essencial para obter nosso objetivo é o MCT, explicado na Subseção 2.3.2. Utilizamos uma simulação microscópica, sendo que a distinção entre macro e micro é descrita na Subseção 2.3.3.

2.3.1 Fluxo

Veículos devem ser alocados em rotas especificas de forma a minimizar o que chamamos de "fluxo"[1], que expressa o quão fácil ou dificilmente o tráfego move-se através de um curso e aumenta à medida que mais veículos trafegam por ele. O fluxo é equivalente ao tempo de viagem em um percurso. Essa propriedade do fluxo significa que normalmente não basta todos os veículos se dirigirem pelas rotas de menor distância porque isso as deixará congestionadas, i.e. aumentará o fluxo de forma a prejudicar o tempo de viagem.

Há diversas funções determinadas a partir de observações empíricas capazes de expressar a relação entre a quantidade de veículos em uma rua e o seu fluxo, sendo a mais famosa a do Departamento de Estradas Públicas dos Estados Unidos, também conhecida

[1] Também chamado na literatura por "latência".

como função BPR a partir da sigla em inglês (Bureau of Public Roads, 1964). Essas funções são chamadas de funções de fluxo. Neste trabalho não nos limitamos a apenas uma ou outra função de fluxo em particular, mas elas estão sujeitas às restrições especificadas na Subseção 2.5.2.

2.3.2 MCT

A tarifagem de custo marginal (*"marginal-cost tolling"* – MCT) é uma abordagem que busca convergir o UE ao SO fazendo com que cada agente seja cobrado proporcionalmente ao custo que impõe aos demais (Pigou, 1920), assim minimizando o PoA. É um custo de pedágio que é imposto aos veículos transeuntes em adição ao custo de simplesmente atravessar a via, influenciando nas decisões dos agentes já que todos eles buscam minimizar seus próprios custos.

Beckmann et al. (1957) demonstram que o MCT pode ser o suficiente para garantir a convergência (i.e. garantir que o PoA chegue a 1) se for equivalente ao produto da quantidade de veículos pela derivada da função de fluxo de uma via. Em outras palavras, seja x_l a quantidade de veículos passando por uma via l e $f_l(x_l)$ a função de fluxo da via l, o MCT ser $x_l \cdot f_l'(x_l)$ ou equivalente é o suficiente para haver convergência.

2.3.3 Simulações macro e microscópicas

Nosso sistema é uma simulação de fluxo de tráfego *macroscópica*, o que significa que determinamos a quantidade de fluxo nos elos do sistema sem precisar saber em qual momento específico cada veículo passou pelo elo. Em essência, o fluxo é quantificado através de equações diferenciais *parciais* e o cálculo é feito a nível de sistema. Em contrapartida, numa simulação *microscópica* o cálculo é feito a nível de veículo (a partir da velocidade e posição de cada veículo), se valendo de equações diferenciais *ordinárias*. Elas estão relacionadas na medida em que as equações de simulações macroscópicas são as integrais das equações de simulações microscópicas (Francesco and Rosini, 2015).

2.4 Aprendizagem por reforço

Na aprendizagem por reforço um agente aprende por tentativa e erro como se comportar em um dado ambiente (Joshi et al., 1996; Sutton and Barto, 1998). De maneira geral ele funciona da seguinte forma: um agente dotado de conhecimento observa o estado atual do seu ambiente e escolhe uma ação baseado no que conhece. Ao executar a ação o agente é recompensado e ele usa essa recompensa para atualizar seu conhecimento. Um ciclo completo de aprendizagem por reforço é chamado de um "episódio". Diversos episódios podem ser realizados em sucessão até se chegar num nível de conhecimento satisfatório.

Um problema de aprendizagem por reforço pode ser formulado como um "processo de decisão de Markov" (*Markov decision process* – MDP): uma quádrupla (S, A, T, r) onde S representa o conjunto de possíveis estados, A o conjunto de possíveis ações, $T : S \times A \times S \to [0, 1]$ uma função de transição, e $r : S \times A \to \mathbb{R}$ a função de recompensa.

Um método apropriado para nossos agentes aprenderem qual a melhor rota escolher é o algoritmo de aprendizagem-Q ("Q-*learning*"), que veremos em seguida.

2.4.1 Q-*learning*

O Q-*learning* (cujo "Q" significa "**Q**ualidade") é um algoritmo de aprendizagem por reforço independente de modelos e desenvolvido por Watkins (1989). O Q-*learning* é baseado na exploração com tentativa e erro a fim de computar uma função $Q(s, a)$, que retorna a recompensa estimada por se realizar a ação a no estado s, com o domínio $Q : S \times A \to \mathbb{R}$. Ele tem a garantia de convergir a valores ótimos se todos os pares de estado-ação forem experimentados infinitas vezes em um sistema de um único agente (Watkins and Dayan, 1992).

Na prática, Q é uma tabela de valores guardada em memória que precisa ser atualizada durante o processo de aprendizagem, podendo ser chamada de "tabela-Q". A cada passo t tomado pelo agente no estado s e escolhendo uma ação a ele receberá uma recompensa $r_t(s_t, a_t)$, que será usada para computar um novo valor de $Q(s, a)$ da seguinte forma:

$$\overbrace{Q_t(s_t, a_t)}^{\text{novo}} \leftarrow (1 - \alpha) \cdot \overbrace{Q_{t-1}(s_{t-1}, a_{t-1})}^{\text{antigo}} + \alpha \cdot r_t(s_t, a_t) \tag{2.2}$$

Atualizando, então, a tabela na memória.

Os valores antigos vão decrescendo ao longo do tempo, ou seja, recompensas antigas vão perdendo relevância, como se fosse um gradual "esquecimento", para permitir que novas recompensas tenham peso maior. O quão rápido ocorre esse "esquecimento" é determinado pelo valor $\alpha \in (0, 1]$ que representa a taxa de aprendizagem, ou seja, o quanto o valor antigo deve ser retido comparado ao novo (quanto maior o α menor o peso dos valores antigos).

2.4.1.1 Exploração por ϵ-guloso

Diversas heurísticas podem ser usadas para determinar como o agente escolhe sua ação. Numa heurística gulosa simples o agente sempre escolhe a ação com maior recompensa registrada na tabela Q até então. Em outras palavras, seja a_t a ação do momento t e A o conjunto de ações disponíveis (ignorando estados, para fim de simplificação), a escolha de uma ação seria expressa por: $a_t \leftarrow \text{argmax}_A \, Q_t$.

Porém, para impedir que os agentes fiquem presos a um máximo local e assim garantir que todas as ações sejam experimentadas, usa-se a exploração pelo método "ϵ-guloso" (ϵ/epsilon-*greedy*). Nesse método um fator $\epsilon \in [0, 1]$ define a probabilidade do agente escolher uma ação aleatória ao invés da ação com maior recompensa que ele encontrou até aquele instante (Sutton and Barto, 1998). A escolha de uma ação a_t em um dado momento t é denotada na Equação 2.3.

$$a_t \leftarrow \begin{cases} \mathcal{U}_A, & \text{se } \mathcal{U}_{[0,1)} < \epsilon; \\ \operatorname{argmax}_A Q_t, & \text{caso contrário.} \end{cases} \tag{2.3}$$

2.5 TQ-*learning* (*baseline*)

A "aprendizagem-Q baseada em pedágios" (*Toll-based Q-learning* – TQ-*learning*) é um aprimoramento do Q-*learning* desenvolvido por Ramos et al. (2020a) para lidar com sistemas multiagentes e, através da aplicação de MCT, minimizar o PoA, fazendo com que o desempenho de uma rede de tráfego convirja ao valor de seu SO ao invés do valor de seu UE. Ele propõe-se a ser uma simulação macroscópica[2] de tráfego em que os motoristas estão sujeitos a aprendizagem por reforço.

Em uma dada malha de tráfego representando as ruas de uma cidade, são distribuídos motoristas que querem encontrar os trajetos com o menor custo para irem de uma origem a um destino fixados para cada motorista. O custo é o tempo de viagem somado a um possível pedágio associado ao trajeto. No sistema podem ocorrer congestionamentos: quanto mais veículos estão passando por uma rua, maior tende a ser o tempo de viagem.

Motoristas buscam o menor custo e podem evitar trajetos onde o custo começa a se tornar elevado. Seguindo o princípio do MCT, pedágio é determinado de forma a ser proporcional ao quanto de prejuízo os motoristas causam aos demais por trafegar em uma rota específica – ou seja, ao quanto sua presença causou aumento no tempo de viagem dos demais. Sendo obrigatório todos os motoristas pagarem e supondo que todos os motoristas buscarão minimizar custos, a tendência é que o tempo médio de viagem dos motoristas no sistema se torne a mínima possível.

O TQ-*learning* será o *baseline* do nosso trabalho, por isso daremos um tratamento mais aprofundado do seu funcionamento nesta seção, onde descreveremos adiante tudo que dele for essencial para o entendimento de nosso trabalho. Explicamos de que forma ele difere do Q-*learning* tradicional na subseção 2.5.1, detalhamos os conceitos e fórmulas necessários para seu entendimento na subseção 2.5.2 e delineamos seu algoritmo na subseção 2.5.3. Onde for necessário para fins de exposição, apresentaremos formalizações de maneira diferente mas equivalentes às do trabalho original, deixando explícito quais conceitos foram introduzidos pela primeira vez aqui.

[2] Ver subseção 2.3.3 para a distinção entre simulações micro e macro.

2.5.1 Especifidades em relação ao Q-*learning* básico

O TQ-*learning* de Ramos et al. (2020a) difere do Q-*learning* básico em alguns pontos específicos que descreveremos nesta subseção.

O agente terá apenas um estado inicial que não poderá ser mudado, porque em cada episódio ele realiza uma escolha de rota e há a garantia que ele atingirá seu destino. Isso significa que na Equação 2.2 e em todas as demais manipulações das tabelas Q o parâmetro s poderá ser simplesmente ignorado.

A garantia de convergência do Q-*learning* a valores ótimos como mencionado na subseção 2.4.1 aplica-se apenas a sistemas de um único agente mas não necessariamente para sistemas multiagentes (necessariamente não se aplica quando todos atualizam uma mesma tabela-Q). Para resolver esse problema, Ramos et al. (2020a) faz com que cada agente trabalhe com sua própria tabela-Q, possibilitando o aprendizado independente de cada um. As tabelas no TQ-*learning*, portanto, variam não apenas ao longo de instantes t como também ao longo de agentes i, sendo denotada por $Q_{i,t}$.

Este algoritmo usa a exploração por ϵ-guloso mencionada na Subsubseção 2.4.1.1, introduzindo além do fator α também o fator ϵ. Esses dois fatores são sujeitados a uma "deterioração" que é descrita com mais detalhes logo a seguir.

2.5.1.1 Diminuição sistemática de taxas

As taxas de aprendizado α e de exploração ϵ são, a partir dos seus valores iniciais, diminuídas sistematicamente ao longo da aprendizagem até os agentes convergirem a um ponto fixo representando o valor de SO. Isso significa que à medida que o tempo avança o sistema vai se estabilizando, ou seja, se torna cada vez menos randômico e mais determinístico. O comportamento dos agentes vai se "enrigecendo", dando cada vez menos importância para variações de terreno.

A diminuição é determinada por taxas de decaimento. As taxas de decaimento λ e μ regem α e ϵ, respectivamente. Em um dado episódio t, $\alpha(t) = \lambda^t$ e $\epsilon(t) = \mu^t$. Todas as taxas têm valor $\in (0, 1]$.

2.5.2 Conceitos e equações

O sistema busca resolver um problema de tráfego que pode ser definido como a quádrupla $P = (G, D, f, \tau)$. Um grafo $G = (N, L)$ representa a estrutura de uma malha de tráfego, formado por "nós" N que são conectados entre si por "elos" unidirecionais L, representando ruas e interseções. D é o conjunto de motoristas, nossos agentes de Q-*learning*, sendo $d = |D|$ a quantidade total de motoristas. f é o conjunto de equações de fluxo e τ é o conjunto de equações de pedágio.

Uma tupla de dois nós constituindo um nó de "origem" e um de "destino" será referida como par "OD". Um conjunto de elos formando um caminho que conecta os dois nós de um par OD, saindo da origem e terminando no destino, será definido como uma "rota", podendo haver múltiplas rotas para um mesmo par OD. As rotas disponíveis para cada par OD são determinadas antes da execução do problema através do algoritmo KSP[3] de Yen (1971), que identifica as K rotas mais curtas da origem ao destino dentro da rede para cada par OD. K é usado como hiperparâmetro para controlar a quantidade de opções de escolha dos motoristas.

Cada motorista $i \in D$ é um agente de Q-*learning* cujo objetivo de maximização de recompensa consiste em minimizar seus custos de viagem na malha de tráfego, e possui:

1. Um par OD cuja origem representa seu ponto de partida e o destino aonde ele quer chegar, que permanece o mesmo durante toda a existência do motorista ao longo do experimento.
2. Um conjunto A_i contendo as K rotas mais curtas formando caminho entre o par OD de i, que também permanece o mesmo ao longo do experimento. A ação do motorista consiste em escolher uma rota dentre essas alternativas.
3. Em cada episódio t uma ação $a_{i,t} \in A_i$ que representa a rota dentre A_i que o motorista i escolheu. A escolha é realizada usando a Equação 2.3 do método ϵ-guloso.

Cada elo $l \in L$ possui:

1. A quantidade x_l de veículos trafegando nele (também chamada de volume de tráfego), que é atualizado a cada episódio;
2. Uma equação de fluxo $f_l : x_l \to \mathbb{R}^+$ que determina o tempo gasto nesse elo em função da quantidade de veículos x_l;
3. E uma equação de pedágio $\tau_l : x_l \to \mathbb{R}^+$.

A quantidade x_l de veículos no elo é atualizada com base nas escolhas dos motoristas a cada episódio t ao longo do experimento. Embora as fórmulas de f_l e t_l permaneçam sempre as mesmas, os resultados de suas equações variam junto com x_l por serem em função deste. As duas equações são essenciais para se calcular as equações de custo, que por sua vez são usadas para definir a equação de recompensa.

As equações de fluxo f não precisam ser as mesmas para cada l e variam de acordo com a rede, basta que satisfaçam as seguintes condições estabelecidas por Ramos

[3] *k-shortest path routing* – roteamento dos k caminhos mais curtos.

et al. (2020b): devem ser *a*) polinomiais homogêneos, *b*) não-negativas, *c*) diferenciáveis e *d*) univariadas[4]. Isso permite que calculemos suas derivadas para utilizá-las na equação de pedágios a fim de satisfazer a condição de convergência do MCT encontrada por Beckmann et al. (1957)[5].

Definimos a equação de pedágio τ_l na Equação 2.4, onde f'_l é a derivada de f_l. A equação de pedágio realiza o papel de MCT no sistema.

$$\tau_l = x_l \cdot f'_l(x_l) \tag{2.4}$$

Tendo f_l e τ_l podemos definir as equações de custo. A equação de custo básica $c_l : x_l \to \mathbb{R}^+$ para um elo é a soma do seu fluxo ao pedágio:

$$c_l(x_l) = f_l(x_l) + \tau_l(x_l) \tag{2.5}$$

A equação de custo C_R para uma rota R é o somatório dos custos de cada elo l em R:

$$C_R = \sum_{l \in R} c_l \tag{2.6}$$

Lembremos que a ação $a_{i,t}$ escolhida pelo motorista i é na verdade a *rota* que ele vai tomar no episódio t. Sendo assim, a ação pode ser usada na equação de custo. A recompensa da ação é o negativo do custo da rota em que se constitui a ação, fazendo com que quanto menor o custo maior a recompensa. Obtemos dessa forma a seguinte equação de recompensa:

$$r(a_{i,t}) = -C_{a_{i,t}} \tag{2.7}$$

2.5.3 Algoritmo

Munidos dos conceitos principais do TQ-*learning* podemos partir para uma descrição do seu algoritmo. Formalizamos no Algoritmo 1 o passo-a-passo do processo aqui descrito.

O primeiro passo é inicializar a tabela-Q com valores zero pois nada ainda foi explorado. Em seguida, executa-se $T \in \mathbb{N}$ episódios de Q-*learning*. Cada episódio t é constituído de três etapas:

1. Os fatores α e ϵ são atualizados, e o conjunto x de volumes de tráfego é (re)inicializado com zero para cada elo l. Cada motorista i escolhe uma rota $a_{i,t}$ como ação de acordo com a Equação 2.3, e x_l é incrementado para cada elo l na rota.

[4] Mais informações em (Ramos et al., 2020a, p. 8)
[5] Ver subseção 2.3.2

2. Tendo sido distribuídos todos os motoristas i o conjunto x se encontra com valores atualizados e podemos recalcular as recompensas a partir dos novos volumes em x e as equações em f e τ. Atualiza-se a tabela-Q de cada motorista i recompensando-os com base na Equação 2.7.

3. Calcula-se o tempo médio de viagem para o episódio t. Para fins expositivos, neste trabalho denotamos esse valor como w_t e o formalizamos na Equação 2.8. Ela é o somatório dos tempos de viagem de todos os motoristas naquele episódio, dividido pela quantidade d de motoristas. O tempo de viagem de um motorista i específico em um dado episódio t é a soma dos tempos de viagem para cada elo na sua rota escolhida $a_{i,t}$.

$$w_t \leftarrow \frac{1}{d} \cdot \sum_{i \in D} \sum_{l \in a_{i,t}} f_l(x_l) \tag{2.8}$$

Processados todos os T episódios, ao final do processo temos uma tabela-Q otimizada e o conjunto de tempos médios de viagem w, obtidos para cada episódio t. Conclui-se assim a execução do algoritmo. O tempo médio de viagem final é denotado por v, e é obtido do último episódio: $v = w_T$. O valor v será usado para avaliação desempenho ao longo deste trabalho.

2.6 Comparação de proximidade

Como vimos na subseção 2.5.3, o tempo médio de viagem final encontrado na execução do algoritmo é denotado por v. Esse valor pode ser comparado com outro, um valor de referência denotado por v^*, para determinar o quanto v está próximo dessa referência v^*. Esse cálculo de proximidade é feito pela fórmula de $\phi : v \times v^* \rightarrow \mathbb{R}$ na Equação 2.9 (Ramos et al., 2020a, p. 16), sendo que $\phi \leqslant 1$. Quanto mais próximo de 1 for o valor de ϕ, mais próximo v está de v^*. Na prática, ϕ é usado para avaliar o quanto o resultado de um dado experimento se aproximou de valores de UE e SO ideais, calculados de antemão para cada rede e obtidos da literatura[6].

$$\phi(v; v^*) = 1 - \frac{|v - v^*|}{v^*} \tag{2.9}$$

Munidos dos conceitos mais importantes para a compreensão dos trabalhos relacionados e do nosso método proposto, podemos apresentá-los nos capítulos seguintes com o embasamento apropriado.

[6] Ver Subseção 5.2 para os valores de referência e suas fontes.

Algoritmo 1 Q-*learning* baseado em pedágios

function TQ-LEARNING($P; T; K; \lambda; \mu$)
  $(G, D, f, \tau) := P$;
  $(N, L) := G$;
  $A \leftarrow$ inicialização de opções de rotas em função de K;

  ▷ Inicialização da tabela-Q
  ▷ Começamos em $t = 0$ aqui para, no 1º episódio ($t = 1$), $Q_{i,t-1}$ ser zero
  for $t \in [0..T]$ **do**
    $Q_{i,t}(a) \leftarrow 0 \ \forall i \in D, \forall a \in A_i$;

  ▷ Para cada episódio t...
  for $t \in [1..T]$ **do**
    ▷ 1ª etapa
    $\alpha \leftarrow \lambda^t$; ▷ atualiza fator de aprendizado pro episódio
    $\epsilon \leftarrow \mu^t$; ▷ atualiza fator de exploração pro episódio
    $x_l \leftarrow 0 \ \forall l \in L$; ▷ reinicia-se os volumes de tráfego de todos os elos l
    ▷ Motoristas escolhem rotas e os volumes de tráfego são atualizados
    for $i \in D$ **do**
      $a_{i,t} \leftarrow$ Equação 2.3; ▷ motorista escolhe uma rota como ação
      $\forall l \in a_{i,t} \mid x_l \leftarrow x_l + 1$; ▷ incrementa-se o x_l de cada elo l na rota $a_{i,t}$

    ▷ 2ª etapa
    ▷ f, τ, c, C e r retornarão novos resultados em função dos novos valores de x
    for $i \in D$ **do**
      ▷ Atualiza o valor na tabela-Q com base no novo cálculo de r
      $Q_{i,t}(a_{i,t}) \leftarrow (1 - \alpha) \cdot Q_{i,t-1}(a_{i,t-1}) + \alpha \cdot r(a_{i,t})$;

    ▷ 3ª etapa
    $w_t \leftarrow$ Equação 2.8 ▷ tempo médio de viagem dos veículos no episódio

  return w_T ▷ tempo médio de viagem final: v

3 Trabalhos relacionados

Neste capítulo falamos sobre trabalhos relacionados encontrados na literatura, descrevendo as tecnologias utilizadas e como elas se alinham ou diferem dos nossos objetivos. Na Seção 3.1 apresentamos variantes do algoritmo de otimização por otimização de formigas, uma abordagem notável por ser altamente descentralizada e já ter sido utilizada para obter aproximações precisas. Na Seção 3.2 discutimos sobre abordagens com tarifagem *a priori*, significando que os valores dos pedágios a serem cobrados são calculados antes do motorista escolher sua rota. Os trabalhos na Seção 3.3 são variantes do Q-*learning* e usam tarifagem *a posteriori*, sendo um deles o nosso *baseline*. Por fim, na Seção 3.4 mencionamos os trabalhos que não se enquadram apenas em uma das categorias mencionadas anteriormente e então finalizamos o capítulo resumindo o diferencial do nosso trabalho.

3.1 Algoritmos de otimização por colônias de formigas

Os algoritmos de otimização por colônia de formigas (*"ant colony optimization"* – ACO) são uma família de algoritmos que se inspiram no comportamento de formigas reais ao navegar pelo ambiente para implementar agentes virtuais, com o objetivo geral de encontrar caminhos mais curtos em problemas de grafos (Dorigo et al., 1991). Formigas reais conseguem usam feromônios com um efeito de atração para coordenar a busca por recursos com outras formigas da colônia. Inspirado nisso, "formigas" virtuais (i.e. agentes simulados) se locomovem por um espaço registrando suas posições e a qualidade das soluções que encontraram até então, a fim de que no decorrer da simulação outros agentes possam usar as mesmas informações para continuar aprimorando a solução.

ACO's foram inicialmente desenvolvidos para obter soluções aproximadas de caminhos mais curtos, mas são também usados em diversos trabalhos para investigar o problema de distribuição de tráfego. Eles demonstram a possibilidade de se usar sistemas multiagentes e aprendizagem por reforço para encontrar equilíbrios de sistema em simulações de tráfego. No trabalho de D'Acierno et al. (2006), por exemplo, o algoritmo é adaptado de forma que as formigas convirjam a um comportamento equivalente ao de motoristas em UE. Porém, ele não almeja a obtenção do SO.

O ACO foi adaptado por Dias et al. (2014) para a diminuição de congestionamentos essencialmente invertendo-se a lógica dos feromônios, de forma que os agentes são repelidos pelos feromônios ao invés de atraídos, e assim diminuindo congestionamentos. Essa abordagem é chamada pelo autor de IACO, ou *Inverted Ant Colony Optimization* ("otimização por colônia de formigas invertida"). Agentes usando apenas IACO não conseguiriam encontrar o caminho para seus destinos, então o autor precisou combinar a

abordagem com o algoritmo de Djikstra, que obtém os caminhos espacialmente mais curtos e assim eles servem para os agentes como uma heurística de por onde seguir enquanto evitam feromônios.

A abordagem resultante melhora significativamente não só o desempenho do sistema como também o consumo de combustível dos motoristas. Ainda por cima ela permite o controle da proporção de usuários que aderem ou não ao sistema de feromônios, produzindo resultados intermediários que permitem a análise da possível adesão gradual do método.

Apesar das vantagens do trabalho de Dias et al. (2014), os autores não oferecem garantias de que o sistema convirja ao SO. Veremos adiante diversas técnicas que têm essa garantia. Além disso, sua dependência no algoritmo de Djikstra introduz um fator de centralização e complexidade pelo cálculo exigir conhecimento de todo o sistema para produzir os caminhos mais curtos. Essa dependência indica que o IACO não é suficiente para a resolução do problema, sendo sua eficácia condicionada por modelos adjacentes. Outro problema é que, no tipo de situação real contemplada pelo trabalho, o análogo a "feromônios" seriam informações armazenadas e distribuídas por um agente centralizador.

No trabalho de Jabbarpour et al. (2014), congestionamentos são previstos no curto prazo e subsequentemente evitados usando ACO. A qualidade do algoritmo depende fortemente do método utilizado para as previsões, que podem ser redes neurais ou outras formas de aprendizagem de máquina. Isso introduz uma certa flexibilidade, mas novamente a abordagem ACO não é suficiente e depende de modelos adjacentes que podem ser bem mais complexos do que o ACO em si. Além disso, muitos dados precisam ser coletados globalmente tanto para possibilitar as previsões em si quanto para permitir a coordenação entre agentes a partir dos "feromônios" virtuais.

No geral, ACO's são sempre simulações microscópicas[1], mas nosso trabalho será com simulações macroscópicas.

3.2 Abordagens com tarifagem *a priori*

O uso de pedágios para diminuição de congestionamentos destaca-se pela sua simplicidade e parcimônia de pressupostos sobre o modelo, e há pelo menos duas formas de cobrá-los: com valores estabelecidos antes do motorista passar pela rota (o que seria uma abordagem *a priori*) e com valores estabelecidos só depois (ou seja, *a posteriori*). Nesta seção falaremos sobre abordagens *a priori* relevantes para o nosso trabalho.

É possível calcular pedágios fixos que levem ao desempenho ótimo do sistema, inclusive encontrando as menores tarifas necessárias (Hearn and Ramana, 1998). O pro-

[1] Ver Subseção 2.3.3.

blema de se encontrar valores tarifários fixos e ideais para um sistema de tráfego é também conhecido como "problema das cabines de pedágio" (em inglês, *tollbooth problem*), e é classificado como NP-*hard*, o que implica em sua alta complexidade computacional (Stefanello et al., 2017).

Buscando contornar essa complexidade, o trabalho de Stefanello et al. (2017) oferece uma aproximação eficiente. Mas, tanto no seu trabalho quanto no de Hearn and Ramana (1998), o algoritmo é executado uma única vez, de maneira centralizada, e portanto não é tolerante a mudanças (e.g. alterações nos padrões das ruas), exigindo recálculo para qualquer modificação.

Descrito na Subseção 2.3.2, o MCT é uma forma de tornar um sistema de tráfego convergente ao SO fazendo com que os agentes sejam cobrados pelos custos que impõem aos demais por suas ações. Há trabalhos que exploram o potencial do MCT para o gerenciamento dinâmico dos pedágios, como o Δ-*tolling* (Sharon et al., 2017; Mirzaei et al., 2018). Um sistema centralizado ainda precisa coletar informações e determinar os pedágios com base em dados históricos de congestionamento, mas o cálculo dos pedágios passa a ser dinâmico e, dessa forma, mais adaptável.

Porém, o verdadeiro impacto que um motorista causou aos demais só pode ser determinado após o fato (a não ser que alguma previsão seja feita), então motoristas podem acabar pagando pedágios com custos desproporcionais ao impacto que realmente causaram aos demais. Por exemplo, no caso em que os congestionamentos de um dia foram excepcionalmente menores do que no passado. Isso significa que tais sistemas de pedágio podem ser injustos para os motoristas.

Uma abordagem similar ao MCT é o de "recompensas diferenciais" (no inglês original, *difference rewards*), introduzido por Wolpert and Tumer (1999). Aproximações eficazes foram desenvolvidas por Agogino and Tumer (2004) e Colby et al. (2016). Nessa técnica, um agente é recompensado (ou penalizado) por sua ação proporcionalmente ao quanto que o desempenho do sistema melhorou (ou piorou) por conta da ação escolhida. A recompensa diferencial $D_i(a_i)$ concedida a um agente i pela sua escolha a_i após ser determinada pelo cálculo $D_i(a_i) = G(\alpha) - G(\alpha_{-i})$, onde α é a ação conjunta de todos os agentes, α_{-i} é todas as ações exceto a ação a_i, e $G(\cdot)$ é o sinalizador de recompensa global (no caso em que estamos lidando com tráfegos, pode representar o tempo médio de viagem no sistema dado um conjunto de ações).

O problema associado a essa técnica é que, apesar de ser mais justa, ela ainda depende de informações globais (por exemplo, para calcular o valor de $G(\alpha)$) e leva mais episódios para convergir do que o Δ-*tolling* supracitado. Para calcular as recompensas é necessário uma autoridade centralizada capaz de observar todo o sistema, o que não é um pressuposto muito apropriado para situações reais.

O Δ-*tolling* e as recompensas diferenciais ambos trabalham com simulações macroscópicas.

Veremos adiante como que técnicas baseadas em Q-*learning* nos permitem a tarifagem *a posteriori*, trazendo maior descentralização no cálculo de pedágios e tarifas mais justas. Inclusive, as técnicas apresentadas obtêm convergência do resultado em menor quantidade de episódios.

3.3 Aprimoramentos do Q-*learning*

Os trabalhos desta seção adaptam o método de Q-*learning* tradicional para sistemas de tráfegos e multiagentes. O principal, TQ-*learning*, adapta o Q-*learning* para levar a convergência do sistema ao SO através de pedágios e será usado como *baseline* para o nosso projeto. Ele é aprimorado pelo GTQ-*learning*, que permite que agentes tenham preferências heterogêneas através da introdução de um fator η e ainda mantendo a convergência do sistema ao SO. Todas as técnicas apresentadas nesta seção são simulações macroscópicas.

3.3.1 TQ-*learning*

O TQ-*learning* apresenta uma abordagem inteiramente descentralizada de tarifagem *a posteriori*, demonstrando convergência em menos episódios do que nas supracitadas obras com tarifagem *a priori* (Ramos et al., 2020a). Ele adapta o Q-*learning* para um ambiente de múltiplos agentes e usa o MCT para garantir a convergência ao SO no comportamento interativo entre os agentes.

A técnica MCT não requere conhecimento global do sistema porque os cálculos com MCT se baseiam em dados locais referentes apenas aos trajetos pelos quais os motoristas passaram. Sem necessidade de gerenciamento centralizado, os próprios motoristas se auto-corrigem buscando diminuir o valor gasto em pedágios. Sendo os pedágios sempre proporcionais ao prejuízo que o motorista causa aos demais, a auto-correção de cada um independentemente naturalmente leva a melhoras no sistema como um todo. Isso confere ao TQ-*learning* um alto nível de descentralização.

Para alcançar seus objetivos o TQ-*learning* pressupõe que todos os agentes são participantes do sistema de pedágios, ou seja, que todos sempre pagam. Mas na prática dificilmente poderemos contar com essa garantia, por conta de fatores como a evasão de motoristas ou limitações econômicas. Consideremos, por exemplo, a sugestão do autor de que o pagamento de pedágios seja realizado por um dispositivo instalado no veículo do motorista. Haveria, portanto, um custo na obtenção e instalação desse dispositivo. Se esse dispositivo fosse disponibilizado por uma empresa privada, seria necessário haver incentivo o suficiente para as pessoas voluntariamente se tornarem usuárias do serviço.

Caso a adoção do dispositivo fosse garantida por alguma lei que obrigasse o uso, haveria o custo da própria implementação da lei (além de introduzir um fator de centralização).

De jeito ou de outro, não seria realista esperar uma adesão plena e imediata ao sistema de pedágios, e se faz necessário investigar como a abordagem do TQ-*learning* lidaria com proporções intermediárias de adesão.

3.3.2 GTQ-*learning*

O próprio TQ-*learning* por sua vez é aprimorado através do método de "aprendizagem-Q baseado em pedágios generalizado" (*Generalized Toll-based Q-learning* – GTQ-*learning*). Ele introduz a possibilidade de preferências heterogêneas entre os agentes, ainda mantendo a convergência do sistema ao SO (Ramos et al., 2020b).

Mais especificamente, o cálculo de custo de cada motorista i é determinado por um fator individual $\eta_i \in [0, 1]$ que determina qual a prioridade do usuário em suas viagens: minimizar o tempo (quanto mais próximo η_i for de 0) ou minimizar o valor de pedágio (quanto mais próximo η_i for de 1). A preferência η_i permanece fixa para cada motorista i sem variar com o tempo. O cálculo do pedágio é feito de forma a anular as diferenças de preferência ao longo prazo, levando à mesma convergência ao SO que o TQ-*learning* traz.

Estratégias são desenvolvidas para penalizar usuários cujo comportamento não aparenta ser compatível com sua preferência proferida, o que consiste numa forma de evasão. O GTQ-*learning* já busca uma forma de lidar com motoristas que tentam burlar o sistema mas, assim como o trabalho do TQ-*learning*, ainda pressupõe que todos os motoristas pagarão pedágios, no geral permanecendo sujeito às mesmas ressalvas.

3.4 Outros

Klügl et al. (2021) investigam os efeitos do compartilhamento de informações entre agentes num sistema que busca a aproximação de UE, com resultados demonstrando convergência em ainda menos episódios do que os trabalhos de Q-*learning* supracitados. Porém, ele foca exclusivamente em acelerar a convergência da aproximação de UE, sem verificar a aproximação de SO. Portanto, não há o uso de pedágios. A técnica apresentada parte de premissas realistas e parece ter o potencial de otimizar ainda mais a eficiência do TQ-*learning*, mas essa integração está fora de nosso escopo. A simulação é macroscópica.

Legge (2005) combina Q-*learning*, ACO e redes neurais em um único modelo. Em sistemas com muitos estados e ações a tabela-Q tradicional se torna difícil de manejar, mas é possível usar uma tabela de dimensão reduzida com valores aproximados. A obtenção de valores-Q aproximados para uma combinação específica de estado e ação é feita através de uma rede neural pré-treinada. A função de recompensa incentiva escolhas que minimizem

a mudança nas probabilidades de uma ação específica ser escolhida. O sistema consegue reagir a congestionamentos causados por aumento de tráfego e se estabilizar, mas sua preocupação principal é manter a estabilidade de escolhas ao invés de convergir ao UE ou ao SO.

O presente trabalho apresenta uma nova versão do TQ-*learning*, visto que essa técnica é reconhecidamente mais eficaz e viável para realizar experimentos do que as demais apresentadas. Nossa proposta é uma alteração no algoritmo permitindo que nem todos os motoristas sejam cooperadores do sistema com base em circunstâncias dinâmicas, com o objetivo de analisar mais cenários e aumentar o potencial da ferramenta em situações reais. Satisfazemos, então, todas as seguintes propriedades que não se encontram simultaneamente em qualquer um dos artigos supracitados: a abordagem é descentralizada, usa MCT, pode ser usada para aproximar tanto o UE quanto o SO, usa simulações macroscópicas e permite o controle da proporção de motoristas que aderem ao sistema de pedágios. No próximo capítulo nosso método será apresentado detalhada e formalmente.

4 Método proposto

Neste capítulo descreveremos a técnica de Q-*learning* multiagente para resolução de congestionamentos com sistema de pedágios e controle condicional de pagamentos, ou "Q-*learning* baseado em pedágios com pagamentos circunstanciais". Seu objetivo é ser um algoritmo capaz de simular redes de tráfego e mitigar congestionamentos através de MCT permitindo que nem todos os agentes paguem pedágios, sendo este pagamento controlado por condições.

Nosso método se baseia no TQ-*learning* descrito na Seção 2.5. Nele o pagamento de pedágios é sempre obrigatório para todos os usuários ao fim de cada episódio. Alteramos o algoritmo para que, a depender das circunstâncias de cada motorista em cada episódio, um pagamento de pedágio seja feito ou não. Na nossa implementação estabelecemos que a condição de pagamento é determinada por duas subcondições que, no caso em que pelo menos uma é satisfeita, o motorista deve pagar pedágios; caso contrário, ele não paga. Em nossos experimentos as subcondições nunca se manifestam simultaneamente (para fins de simplificação de análise), apesar de ser possível que mais de uma subcondição se demonstre verdadeira ao mesmo tempo.

Na Seção 4.1 apresentamos de maneira resumida as alterações feitas no algoritmo, introduzindo os parâmetros υ, ρ e m, que controlarão a condição para o pagamento de pedágios. Em seguida, detalhamos formalmente na Seção 4.2 os conceitos necessários para a compreensão do método incluindo as fórmulas utilizadas; nela, falaremos com mais detalhes sobre os novos parâmetros e como eles influenciam nos pedágios. Por fim, explicamos na Seção 4.3 o passo-a-passo do algoritmo.

4.1 As alterações no algoritmo de TQ-*learning*

A mudança essencial no algoritmo TQ-*learning* foi a introdução de uma condição que deve ser calculada para o algoritmo determinar se, em um dado momento, o pedágio deve ser pago ou não. Em teoria, essa condição pode ser qualquer proposição lógica que possa ser expressa como equação booleana. Neste trabalho definimos que a condição seria constituída de duas subcondições específicas de forma que, se pelo menos uma for verdadeira, a condição inteira é verdadeira. Resumidamente, se tomarmos $\mathcal{P}$ como a condição e p_1 e p_2 como as subcondições, temos que $\mathcal{P} = p_1 \vee p_2$.

A primeira subcondição, ou *subcondição de usuário*, refere-se ao motorista ser "usuário do sistema de pedágios" ou não, sendo satisfeita quando ele o é. Daqui pra frente os dois tipos de motoristas serão referidos como apenas "usuários" e "não-usuários",

respectivamente. Resumidamente, motoristas são consignados a serem usuários ou não no começo de cada experimento e o usuário terá pedágios cobrados ao final de cada trajeto que ele realizar. Os não-usuários pagam pedágio apenas se outra subcondição for satisfeita. A distribuição de usuários é controlada pelo parâmetro υ.

A segunda subcondição, ou *subcondição de trajeto*, é determinada pelos elos mais movimentados em um dado momento do sistema. Se o motorista em seu trajeto passar por elos que se encontram entre os mais movimentados a partir de um certo limiar, a condição é satisfeita e ele deve pagar pedágio. Consideramos que um elo é mais movimentado do que outro quando aquele tem maior volume de tráfego do que este. O limiar é determinado pelo parâmetro $\rho \in [0, 1]$, de forma que os elos acima do limiar formam $100\rho\%$ dos elos mais movimentados[1].

Quando a segunda condição é satisfeita investigamos duas formas mutuamente excludentes de se pagar o pedágio. São os *modos de pagamento*, controlados pelo parâmetro m. No primeiro modo ou *modo rota*, o pedágio vale pelo trajeto inteiro – ou seja, se pelo menos um elo estiver acima do limiar de ρ, paga-se o pedágio também dos demais elos. No segundo modo ou *modo elo*, paga-se pedágio apenas para cada elo que realmente ativou a condição.

Veremos na Seção 4.2 as adaptações introduzidas ao TQ-*learning* para que as novas funcionalidades pudessem ser integradas ao algoritmo.

4.2 Conceitos e equações

Sendo nosso método uma expansão das funcionalidades do TQ-*learning*, boa parte dos conceitos já se encontram apresentados na Subseção 2.5.2. Aqui focaremos no que há de novo ou diferente com relação ao TQ-*learning*, revisando brevemente o que for necessário.

A cada motorista $i \in D$ foi adicionada a seguinte propriedade: um booleano $s_i \in \mathbb{B}$ que determina se ele participa do programa de pedágios (caso s_i verdadeiro) ou não (caso $\neg s_i$), sendo nos respectivos casos um "usuário" ou "não-usuário". Esse booleano é um dos fatores que determinam, na equação de custo, se o motorista pagará pedágio ou não. Assim como o par OD, o valor s_i de cada motorista i é o mesmo do começo ao fim do experimento.

A distribuição de motoristas entre usuários e não-usuários em um dado experimento é controlado pelo hiperparâmetro $\upsilon \in [0, 1]$; ele determina que um motorista i tem a probabilidade υ de ser definido como usuário. Consequentemente, a quantidade de motoristas usuários gira em torno de $\upsilon \cdot d$; a quantidade de não-usuários, em torno de

[1] Exemplo: se $\rho = 0.25$, então estamos lidando com 25% dos elos mais movimentados.

$(1 - \upsilon) \cdot d$.

Em outras palavras, υ controla a proporção de motoristas que são usuários, considerando uma distribuição aleatória. Alguns exemplos de seu efeito na prática: quando $\upsilon = 0$, nenhum motorista é usuário; quando $\upsilon = 1$, todos são; quando $\upsilon = 0.5$, aproximadamente metade dos motoristas é constituída de usuários e a outra metade de não-usuários; etc.

O hiperparâmetro $\rho \in [0, 1]$ controla o limiar de volume de tráfego acima do qual elos mais movimentados são considerados passíveis de pedágios. Para determinar esses elos, precisamos obter $100\rho\%$ dos elos mais movimentados, o que pode ser feito da maneira a seguir.

Seja $\varrho \in \mathbb{N}_0$ o produto de ρ por $|L|$ arredondado pra baixo (ver a Equação 4.1), uma variável ς é o conjunto com os ϱ elos mais movimentados de L (caso $\varrho = 0$, então $\varsigma = \varnothing$).

Em cada episódio, após os motoristas escolherem suas ações e antes do cálculo das recompensas, ς é atualizado de acordo com o x_l de cada elo l da seguinte forma: seja ξ um conjunto ordenado com todos os elos de L em que os elos estão ordenados pelos seus respectivos volumes de tráfego do maior ao menor volume (ver Equação 4.2), então ς corresponde aos ϱ primeiros elos de ξ como denotado na Equação 4.3 (consequentemente, $|\varsigma| \leqslant |L|$). Assim, ς torna-se o conjunto com $100\rho\%$ dos elos mais movimentados; e, junto com s, é usado na condição de pagamento de pedágio.

$$\varrho = \lfloor \rho \cdot |L| \rfloor \tag{4.1}$$

$$\xi \leftarrow (l \in L : x_{\xi_n} \geqslant x_{\xi_{n+1}} \forall n \in [1..|L|)) \tag{4.2}$$

$$\varsigma \leftarrow \{l \in \xi : l = \xi_n \text{ com } 1 \leqslant n \leqslant \varrho\} \tag{4.3}$$

A função de custo para um elo (que era originalmente no TQ-*learning* a Equação 2.5) agora tem o formato $c_l : x_l \times \varpi \to \mathbb{R}^+$, sendo $\varpi \in \mathbb{B}$ o booleano que indica se o pedágio deve ser considerado ou não para um dado custo de elo. Isso é importante para conseguirmos que, nas devidas circunstâncias, alguns motoristas paguem pedágio e outros não. O custo de um elo é seu fluxo somado ao pedágio, sendo o custo do pedágio efetivamente zerado caso ϖ seja falso, como é denotado na Equação 4.4. O valor de ϖ é calculado a partir da aplicação de nossa condição de pagamento às circunstâncias em que o elo é atravessado.

$$c_l(\varpi) = f_l(x_l) + [\varpi] \cdot \tau_l(x_l) \tag{4.4}$$

Antes de seguirmos para a exposição da nova equação de custo por rota é essencial explicarmos o cômputo da condição de pagamento, porque aquela depende desta. Como

foi dito anteriormente a condição de pagamento depende de duas subcondições, bastando que uma subcondição seja verdadeira para que a condição inteira seja verdadeira.

As duas subcondições básicas são: 1) o motorista i é usuário (ou seja, s_i é verdadeiro), ou 2) o elo l da rota R se encontra dentro do limite ς de elos mais movimentados. Essas subcondições foram introduzidas na Seção 4.1 como subcondição de usuário e subcondição de trajeto, respectivamente. A subcondição de usuário (formalizada na Equação 4.5 com a assinatura $p_1 : i \to \mathbb{B}$) é simples o suficiente pra dispensar mais explicações, mas a subcondição de trajeto merece ser aprofundada nos parágrafos seguintes por conta de sua complexidade.

A equação de custo por rota pode calcular um valor ϖ diferente para cada elo dentro da sua rota, e isso significa que a equação pode controlar se o pedágio será pago para a rota inteira ou apenas para elos individualmente. Isso é necessário para se fazer valer os diferentes modos de pagamento, também introduzidos na Seção 4.1. O cômputo da subcondição de trajeto deve, portanto, levar em consideração tanto R quanto l, e também deve saber qual o modo de pagamento que se está usando.

Definimos um parâmetro $m \in \{\text{ROTA}, \text{ELO}\}$ para controlar qual o modo de pagamento, podendo alternar entre um modo "rota" ($m = \text{ROTA}$) ou um modo "elo" ($m = \text{ELO}$). No modo rota o pedágio é aplicado à rota inteira. No modo elo o pedágio é aplicado apenas aos elos dentro do limite. Dado uma rota R, um elo $l \in R$ e o modo m de pagamento, a subcondição de trajeto é calculada da seguinte forma: a subcondição será verdadeira no caso $m = \text{ROTA}$ quando houver pelo menos um elo em R que se encontra em ς (independente do elo l específico sendo considerado); no caso $m = \text{ELO}$, será verdadeira se o elo l específico encontrar-se em ς. A formalização da função da subcondição de trajeto $p_2 : R \times l \times m \times \varsigma \to \mathbb{B}$ encontra-se na Equação 4.6.

Podemos então encapsular toda a condição de pagamento em uma única função de condição com assinatura $\mathcal{P} : i \times R \times l \times m \times \varsigma \to \mathbb{B}$, encontrando-se formalizada na Equação 4.7. Ela é usada na Equação 4.8 do custo total de rota, que veremos a seguir.

$$p_1(i) = s_i \tag{4.5}$$

$$p_2(R; l; m; \varsigma) = \begin{cases} (l \in R) \wedge ((R \cap \varsigma) \neq \varnothing), & \text{se } m = \text{ROTA} \\ l \in \varsigma, & \text{se } m = \text{ELO} \end{cases} \tag{4.6}$$

$$\mathcal{P}(i; R; l; m; \varsigma) = p_1(i) \vee p_2(R; l; m; \varsigma) \tag{4.7}$$

O custo total da viagem é pago apenas ao final do trajeto realizado por um motorista i ao longo de uma rota R, e é denotado pela função $C_{i,R} : i \times R \to \mathbb{R}^+$. É o somatório das funções de custo de cada elo l de R, incluindo ou não o valor de pedágio a depender do valor da condição de pagamento, computado para cada elo através da função de condição.

Isso é expresso na Equação 4.8, que substitui a Equação 2.6. Dada a assinatura da função de condição (Equação 4.7), vemos que a condição pode variar de acordo com motorista (i), elo (l) e rota (R), constituindo as circunstâncias que determinam se o pedágio será pago ou não. Obtemos assim um pagamento de pedágios circunstancial.

$$C_{i,R} = \sum_{l \in R} c_l(\mathcal{P}(i; R; l; m; \varsigma)) \tag{4.8}$$

4.3 Algoritmo

Tendo definido todos os conceitos essenciais, podemos descrever nosso algoritmo. Tomando como base o algoritmo do TQ-*learning* descrito na Subseção 2.5.3, novamente focaremos nas divergências e inovações, revisando resumidamente o que for necessário.

O primeiro passo é a inicialização. Novamente a tabela-Q é inicializada com valores zero pois nada ainda foi explorado. O conjunto s é inicializado para cada motorista i determinando se ele será usuário ou não com probabilidade υ.

Em seguida, executa-se $T \in \mathbb{N}$ episódios de Q-*learning*. Cada episódio t é constituído de três etapas:

1. Atualizam-se os fatores de aprendizado (α) e de exploração (ϵ). Os motoristas escolhem suas ações; com nas ações escolhidas (que determinam os elos pelos quais os motoristas passarão) atualiza-se também os volumes de tráfego de elos (x). Aqui não há divergência.

2. Determina-se quais são os elos mais movimentados com base na Equação 4.3. Tendo sido distribuídos todos os motoristas ao longo das rotas, o conjunto x portanto se encontra com valores atualizados e podemos recalcular as recompensas a partir dos novos volumes em x e das equações em f e τ. Atualiza-se a tabela-Q de cada motorista i recompensando-os com base na Equação 2.7 de recompensa. Note-se que, porque a função de custo total C mudou da Equação 2.6 para a 4.8 (que, por sua vez, é afetada pela mudança na função de custo por elo da Equação refeq:basic-link-cost para a 4.4), então também observaremos mudanças na recompensa, que é calculada em função do custo.

3. Calcula-se o tempo médio de viagem w_t para o episódio t com base na Equação 2.8. Aqui novamente não há divergência, e ainda temos que $\nu = w_T$ (o tempo médio de viagem final).

Todo esse processo, cujo passo-a-passo se encontra formalizado no Algoritmo 2, constitui um único experimento.

Algoritmo 2 Q-*learning* baseado em pedágios com pagamento circunstancial

function EXPERIMENTO($P; T; K; \lambda; \mu; \upsilon; \rho; m$)
  $(G, D, f, \tau) := P$;
  $(N, L) := G$;
  $A \leftarrow$ inicialização de opções de rotas em função de K;

  ▷ Inicialização de s e da tabela-Q
  for $i \in D$ **do**
    $s_i \leftarrow \mathcal{U}_{[0,1)} < \upsilon$; ▷ inicializa s_i com probabilidade υ
    ▷ Começamos em $t = 0$ aqui para, no 1º episódio ($t = 1$), $Q_{i,t-1}$ ser zero
    for $t \in [0..T]$ **do**
      $\forall a \in A_i \mid Q_{i,t}(a) \leftarrow 0$; ▷ inicializa a tabela-Q

  ▷ Para cada episódio t...
  for $t \in [1..T]$ **do**
    ▷ 1ª etapa
    $\alpha \leftarrow \lambda^t$; ▷ atualiza fator de aprendizado pro episódio
    $\epsilon \leftarrow \mu^t$; ▷ atualiza fator de exploração pro episódio
    $x_l \leftarrow 0\ \forall l \in L$; ▷ reinicia-se os volumes de tráfego de todos os elos l
    ▷ Motoristas escolhem rotas e os volumes de tráfego são atualizados
    for $i \in D$ **do**
      $a_{i,t} \leftarrow$ Equação 2.3; ▷ motorista escolhe uma rota como ação
      $\forall l \in a_{i,t} \mid x_l \leftarrow x_l + 1$; ▷ incrementa-se o x_l de cada elo l na rota $a_{i,t}$

    ▷ 2ª etapa
    $\xi \leftarrow$ Equação 4.2; ▷ elos ordenados por volume de tráfego
    $\varsigma \leftarrow$ Equação 4.3; ▷ elos mais movimentados atualizados com base em ξ
    ▷ f, τ, c, C e r retornarão novos resultados em função dos novos valores de x
    for $i \in D$ **do**
      ▷ Atualiza o valor na tabela-Q com base no novo cálculo de r
      ▷ Aqui as novas Equações 4.4 (c) e 4.8 (C) alteram o comportamento de r
      $Q_{i,t}(a_{i,t}) \leftarrow (1 - \alpha) \cdot Q_{i,t-1}(a_{i,t-1}) + \alpha \cdot r(a_{i,t})$;

    ▷ 3ª etapa
    $w_t \leftarrow$ Equação 2.8 ▷ tempo médio de viagem dos veículos no episódio

  return w_T ▷ tempo médio de viagem final: v

Finalizamos, assim, a exposição detalhada do método proposto: uma variante do TQ-*learning* que permite o pagamento circunstancial de pedágios, as condições específicas utilizadas e os parâmetros de controle que as determinam. Três cenários podem ser obtidos a partir das condições implementadas: o controle de pedágios por usuários, o controle por elos mais movimentados no modo rota de pagamento, e o controle por elos mais movimentos no modo elo de pagamento. Descrevemos o que nosso modelo altera no TQ-*learning*, quais funcionalidades foram adicionadas para atingir nossos objetivos, e

apresentamos o algoritmo de execução do modelo. Seguiremos adiante detalhando os experimentos aplicados ao nosso método, apresentando e analisando os resultados obtidos.

5 Experimentos e resultados

Neste capítulo descrevemos como fizemos nossos experimentos e apresentamos detalhadamente os resultados obtidos. Na Seção 5.1 explicamos o ambiente de experimentação, descrevendo a base de dados de utilizada, os parâmetros escolhidos, a organização entre diversos casos de experimentação e o processo de execução de uma bateria de experimentos com obtenção de resultados. Na Seção 5.2 comparamos os resultados de aproximações do UE e do SO em nossos experimentos com valores de referência obtidos da literatura e nosso *baseline*. Nas seções seguintes apresentamos os resultados para casos com valores intermediários dos parâmetros υ e ρ. Na Seção 5.3 apresentamos os resultados obtidos para os valores intermediários de υ especificamente. Na Seção 5.4 apresentamos os resultados para variações de ρ, que possui dois modos de pagamento.

5.1 Base de dados e execução dos experimentos

Os experimentos foram realizados usando uma seleção das mesmas malhas de tráfego de Ramos et al. (2020a), que será nosso *baseline*. Usamos as mesmas redes exceto a denominada "Sioux-Falls", que leva tempo demais para ser computada mesmo usando quantidades menores de episódios. As redes podem ser divididas em dois grupos: redes obtidas a partir de dados reais (Anaheim e Eastern-Massachusetts) e redes sintéticas (todas as demais). Elas são descritas com mais detalhes a seguir[1].

- $\boldsymbol{B^1, \ldots, B^7}$**:** expansões da rede original usada para ilustrar o paradoxo de Braess. Cada grafo B^p tem $|N| = 2p + 2$ nós, $|L| = 4p + 1$ elos, um par OD, e 4 200 motoristas.

- $\boldsymbol{BB^1, BB^3, BB^5, BB^7}$**:** também expansões das redes de Braess mas com dois pares OD. Cada grafo BB^p tem $|N| = 2p + 6$ nós, $|L| = 4p + 4$ elos e também 4 200 motoristas.

- **OW:** rede sintética de Ortúzar and Willumsen (2011) com $|N| = 13$ nós, $|L| = 48$ elos, 4 pares OD, 1 700 motoristas e rotas que se interseccionam.

- **Anaheim:** abstração da cidade de Anaheim, EUA, com $|N| = 416$ nós, $|L| = 914$ elos, 28 pares OD, 104 694 motoristas e rotas com bastante interseção.

[1] As malhas de tráfego podem ser obtidas do repositório público disponível em https://github.com/goramos/transportation_networks

- **Eastern-Massachusetts (ou EM):** abstração da região leste do estado americano de Massachusetts, com $|N| = 74$ nós, $|L| = 258$ elos, 74 pares OD e 65 576 motoristas. Novamente há bastante interseção entre as rotas.

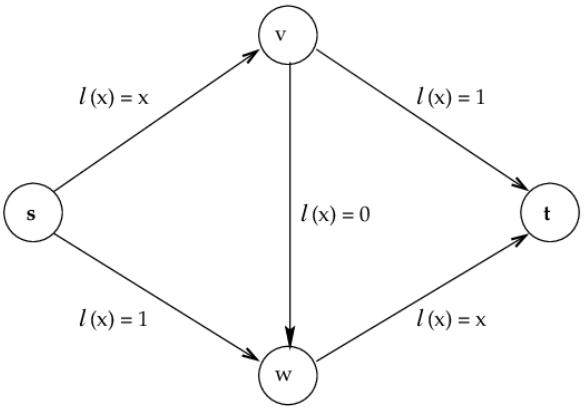

Figura 1: Topologia da rede original do paradoxo de Braess (Lin et al., 2011).

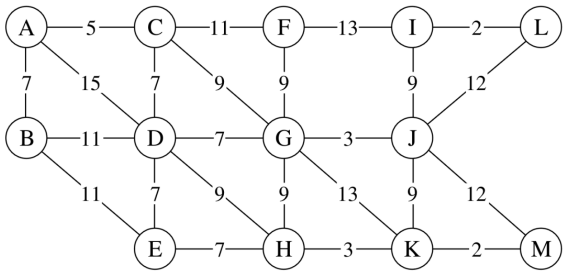

Figura 2: Topologia da rede OW (Ramos and Bazzan, 2015).

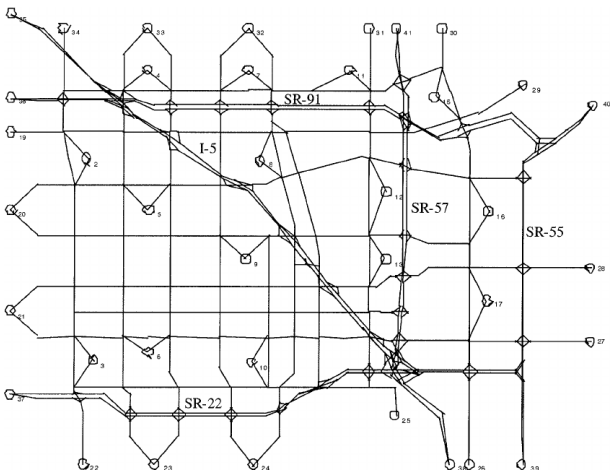

Figura 3: Topologia da rede Anaheim (Sharon et al., 2019).

Cada malha de tráfego não apenas é formada por um grafo G como também é acompanhada das suas próprias equações de fluxo (f) e motoristas (D). Ou seja, uma malha pode ser definida como sendo $\gamma = (G, D, f)$. As equações de pedágio τ de uma

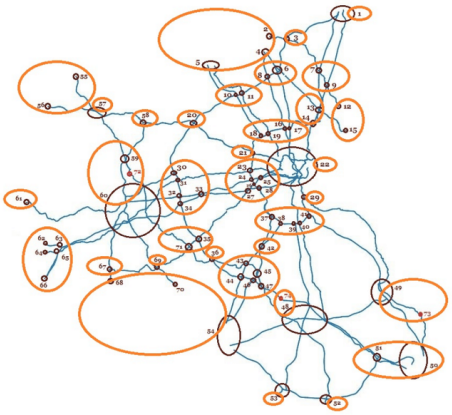

Figura 4: Topologia da rede Eastern-Massachusetts (Sharon et al., 2019).

dada malha podem ser determinadas inteiramente a partir da derivada de cada função em f, o que é feito proceduralmente, e assim nos permite obter um problema $P = (G, D, f, \tau)$ completo para um experimento.

Um experimento completo consiste em repetir uma quantidade T de episódios para uma malha de tráfego G, com os seguintes parâmetros: uma proporção υ de motoristas usuários, uma proporção ρ dos elos mais movimentados que tornam o pedágio obrigatório ao se passar por eles, um valor K que determina a quantidade de rotas a serem disponibilizadas para escolha a cada usuário, e taxas de decaimento λ e μ que determinam os fatores de aprendizado α e de exploração ϵ respectivamente[2]. O algoritmo do experimento foi descrito na Seção 4.3 como o Algoritmo 2.

Rede	K	λ	μ
B^1	3	0.99	0.99
B^2	5	0.99	0.99
B^3	7	0.99	0.99
B^4	9	0.99	0.99
B^5	11	0.99	0.99
B^6	13	0.99	0.99
B^7	15	0.99	0.99
BB^1	3	0.98	0.98
BB^3	8	0.99	0.99
BB^5	4	0.99	0.99
BB^7	4	0.99	0.99
OW	8	0.99	0.99
Anaheim	16	0.995	0.995
Eastern-Massachusetts	16	0.995	0.995

Tabela 1: Parâmetros usados para cada rede durante os experimentos.

Seguindo Ramos (2018), T foi fixado no valor 1 000 para todos os experimentos a

[2] Cf. Subsubseção 2.5.1.1 para mais detalhes sobre o decaimento de fatores.

fim de que rodem em tempo viável. Os parâmetros K, λ e μ são definidos separadamente para cada rede de acordo com a Tabela 1. Exceto pelas redes Anaheim e EM, os parâmetros das demais redes foram obtidos de (Ramos, 2018, p. 113). A determinação dos parâmetros para as redes Anaheim e EM deu-se como a seguir.

As redes Anaheim e EM foram investigadas em Ramos et al. (2020a) com experimentos de 10 000 episódios. Por ser uma quantidade de episódios que resultaria em computações muito demoradas para nossa máquina, diminuímos para 1 000 episódios (igualando com a quantidade de episódios das demais redes). Porém, diminuir a quantidade de episódios significa que não poderíamos aproveitar os mesmos valores para λ e μ de Ramos et al. (2020a), que foram otimizados para 10 000 episódios. Usar os mesmos valores de Ramos et al. (2020a) faria com que α e ϵ não decaíssem rápido o suficiente com 1 000 episódios.

O próprio autor do artigo nos sugeriu, em conversa direta, valores para λ e μ que possam ser usados para experimentos com 1 000 episódios no geral. Seguindo o princípio em (Ramos, 2018, p. 72) de que redes maiores tendem a precisar de K maior, indicando que K é determinado mais pela rede do que pelos demais parâmetros, no fim das contas só o valor de K foi preservado como em Ramos et al. (2020a).

Uma bateria de experimentos consiste em múltiplos experimentos ao longo de permutações dos hiperparâmetros G, υ, ρ e m, sendo $\upsilon, \rho \in M = \{0, 0.25, 0.5, 0.75, 1\}$. Para cada permutação de configuração experimental o experimento é repetido $\sigma = 30$ vezes.

Não consideramos todas as permutações de υ e ρ, apenas as que estão ao longo de cada "eixo", incluindo o caso em que $\upsilon = \rho = 0$ (quando em nenhuma viagem o pedágio é obrigatório). Ou seja, variamos υ ao longo de M mantendo $\rho = 0$, e vice-versa. Os casos $\upsilon = 1, \rho = 0$ e $\upsilon = 0, \rho = 1$ são equivalentes a $\upsilon = \rho = 1$ e têm os três o mesmo efeito: em todas as viagens o pedágio é obrigatório; logo, serão tratados como um único caso, sem experimentos separados para cada um mas apenas para $\upsilon = 1, \rho = 0$.[3] Com essas restrições o espaço de exploração fica menos extenso para cada malha. Além disso, elas simplificam nossas análises análises na medida em que se torna desnecessário verificar a interação entre υ e ρ.

Quando $0 < \rho < 1$, variamos m ao longo de seus possíveis valores $\{\text{ROTA}, \text{ELO}\}$, permitindo que testemos os variados modos de pagamento[4]. Nos demais casos o valor de m se torna irrelevante e portanto ele não precisa variar; para fins de consistência usaremos apenas $m = \text{ROTA}$, o que acaba simbolizando o fato de que quando o pedágio é pago nesses demais casos ele é pago pelo trajeto inteiro.

O Algoritmo 3 formaliza o processo de execução da bateria de experimentos.

[3] Dado que $\upsilon = 1$ já garante que o pedágio sempre será pago, preferimos deixar ρ zerado para simplificar as permutações de m.

[4] Ver Seções 4.1 e 4.2 para o funcionamento dos modos de pagamento.

Algoritmo 3 Bateria de experimentos para uma rede γ

function BATERIA($M; T; \sigma; \gamma; K; \lambda; \mu$)
  $(G, D, f) := \gamma$
  for $\upsilon \in M$ **do**
    for $\rho \in M$ **do**
      ▷ Filtragem de permutações de υ e ρ segundo nossos critérios
      if $(\rho = 0 \leqslant \upsilon \leqslant 1) \vee (\upsilon = 0 < \rho < 1)$ **then**
        $\tau \leftarrow \{g' : g \in f\}$
        for _ $\in [1..\sigma]$ **do** ▷ faça σ vezes; "_" significa valor ignorado
          for $m \in \{\text{ROTA}, \text{ELO}\}$ **do**
            ▷ Filtragem de permutações de m
            if $(0 < \rho < 1) \vee m = \text{ROTA}$ **then**
              $\nu \leftarrow$ EXPERIMENTO$((G, D, f, \tau); T; K; \lambda; \mu; \upsilon; \rho; m)$
              ▷ ν e os parâmetros do experimento são armazenados em arquivo para a agregação de resultados e cálculo de $\overline{\nu}$.

Chamaremos os resultados obtidos no caso em que nenhum motorista paga pedágio ($\upsilon = \rho = 0$) de "aproximação de UE", e no caso em que todos os motoristas pagam pedágio de "aproximação de SO". Os casos em que υ e ρ variam entre 0 e 1 ($0 < \upsilon, \rho < 1$) serão chamados de casos intermediários para o respectivo parâmetro. Valores obtidos em casos intermediários serão por vezes referidos por "valores intermediários".

Denotaremos a média de valores ν obtidos para uma dada configuração experimental por "valor $\overline{\nu}$", também referido como o "desempenho" da rede na configuração específica. O desempenho de uma rede no caso de aproximação de UE será denotado por $\overline{\nu}_{\text{UE}}$, e no de aproximação de SO, por $\overline{\nu}_{\text{SO}}$. Denotaremos por $\Delta\overline{\nu}$ a diferença entre valores $\overline{\nu}$ de casos sucessivos do mesmo parâmetro (υ ou ρ) em uma mesma rede (no caso em que o parâmetro é ρ, também dentro do mesmo modo de pagamento).

5.2 Comparação com experimentos anteriores

As Tabelas 2 e 3 apresentam os resultados encontrados para as aproximações de UE e de SO, respectivamente. Inclui-se também as proximidades entre os valores observados e os de referência, calculadas a partir de ϕ (ver Equação 2.9 na Seção 2.6). O desvio padrão pode ser observado em parênteses. Os valores de referência foram obtidos de (Sharon et al., 2019, p. 6) para as redes Anaheim e EM e, para as demais redes, de (Ramos, 2018, p. 113).

Para as redes sintéticas houve correspondência praticamente exata com as aproximações de SO em (Ramos, 2018, p. 115). No entanto, as aproximações de UE não tiveram uma similaridade tão alta. Salvo três exceções (BB^3, BB^7 e EM), os valores $\overline{\nu}_{\text{UE}}$ observados foram ligeiramente melhores do que os de referência. Isso talvez se deva ao uso de

	UE		
Rede	Referência	Observado ($\overline{v}_{\mathrm{UE}}$)	Proximidade (ϕ)
B^1	20	18.4697 (0.606)	0.9235 (0.030)
B^2	30	27.6516 (1.156)	0.9217 (0.039)
B^3	40	37.3965 (1.138)	0.9349 (0.028)
B^4	50	47.0528 (1.158)	0.9411 (0.023)
B^5	60	56.3216 (0.889)	0.9387 (0.015)
B^6	70	66.2312 (0.682)	0.9462 (0.010)
B^7	80	75.9613 (0.343)	0.9495 (0.004)
BB^1	10	10.0000 (0.000)	1.0000 (0.000)
BB^3	22	22.0091 (0.213)	0.9923 (0.006)
BB^5	50.3	50.5608 (0.039)	0.9948 (10^{-3})
BB^7	123.84	124.1569 (0.140)	0.9974 (0.001)
OW	67.16	67.1986 (0.010)	0.9994 (10^{-4})
Anaheim	13.5625	12.6300 (10^{-3})	0.9312 (10^{-4})
EM	0.4297	0.4362 (10^{-4})	0.9851 (10^{-3})
Média			0.9611 (0.035)

Tabela 2: Valores $\overline{v}_{\mathrm{UE}}$ e suas proximidades aos valores de UE de referência (com desvio padrão).

	SO		
Rede	Referência	Observado ($\overline{v}_{\mathrm{SO}}$)	Proximidade (ϕ)
B^1	15	15.0000 (10^{-4})	1.0000 (10^{-5})
B^2	23.3333	23.3345 (0.004)	1.0000 (10^{-4})
B^3	32.5	32.5004 (10^{-3})	1.0000 (10^{-5})
B^4	42	42.0007 (10^{-3})	1.0000 (10^{-5})
B^5	51.6667	51.6677 (10^{-3})	1.0000 (10^{-5})
B^6	61.43	61.4300 (0.001)	1.0000 (10^{-5})
B^7	71.25	71.2552 (0.006)	0.9999 (10^{-4})
BB^1	7.5	7.5000 (0.000)	1.0000 (0.000)
BB^3	19	19.0001 (10^{-3})	1.0000 (10^{-4})
BB^5	47	47.0033 (0.002)	0.9999 (10^{-4})
BB^7	120.5	120.5410 (0.049)	0.9997 (10^{-3})
OW	66.92	66.9871 (0.006)	0.9990 (10^{-4})
Anaheim	13.3247	12.5475 (10^{-3})	0.9417 (10^{-4})
EM	0.4167	0.4260 (10^{-4})	0.9775 (10^{-3})
Média			0.9941 (0.016)

Tabela 3: Valores $\overline{v}_{\mathrm{SO}}$ e suas proximidades aos valores de SO de referência (com desvio padrão).

parâmetros otimizados no trabalho original para o cálculo da aproximação de SO. Vale ressaltar que a rede BB^1 apresentou aproximações de UE e SO exatamente iguais aos valores de referência, inclusive obtendo variância 0.

As redes Anaheim e EM obtiveram as proximidades mais distantes de 1, mas cada uma por motivos diferentes. A começar pela EM, é possível observar que seus valores observados foram piores do que os de referência tanto para a aproximação de UE quanto a de SO. A rede EM, sendo mais complexa do que as sintéticas, talvez precise de mais episódios para convergir em resultados mais próximos assim como os encontrados por Ramos et al. (2020a) para ela, que utilizou 10 000 episódios ao invés de 1 000.

No caso da Anaheim, a rede demonstrou-se um *outlier*. Seus valores $\overline{v}_{\text{UE}}$ e $\overline{v}_{\text{SO}}$ foram ambos melhores do que o valor de referência do SO. O valor de SO é um limite teórico de desempenho da rede para o melhor desempenho possível, então é esperado que resultados de experimentos nunca o ultrapassem (nem mesmo os de aproximação do SO). No entanto, até mesmo o $\overline{v}_{\text{UE}}$ foi melhor do que o valor de SO de referência. A Anaheim foi a única rede que apresentou tal comportamento. Todavia, veremos adiante que os valores intermediários encontrados para a rede não se destoam dos padrões de outras redes mesmo ao longo dos diferentes controles de pedágio que usamos.

O trabalho de Stefanello et al. (2017), usando a ferramenta CPLEX, calculou um valor de referência para o SO da rede Anaheim que se encaixa com nossas observações: 12.46. Os valores $\overline{v}_{\text{UE}}$ e $\overline{v}_{\text{SO}}$ são ambos maiores do que ele, e a proximidade dele com o $\overline{v}_{\text{SO}}$ seria de aproximadamente 0.9915, que é melhor do que a proximidade com o valor de referência de Sharon et al.. Porém, Stefanello et al. não nos fornece valor de referência para o UE, então, por questão de consistência, mantivemos os valores de Sharon et al. nas tabelas. Talvez os valores encontrados para a rede Anaheim em Sharon et al. (2019) precisem ser revisados, mas isso ficaria para um trabalho futuro.

A rede sintética mais complexa, a OW, teve a menor proximidade dentre as redes sintéticas. Talvez ela também precise, assim como a rede EM, de mais episódios para sua convergência.

As análises de valores intermediários nas próximas seções serão sempre relativas ao $\overline{v}_{\text{UE}}$ e ao $\overline{v}_{\text{SO}}$, não aos valores de referência. Logo, a discrepância das aproximações da rede Anaheim aos seus valores de referência não interfere em nossas análises. É importante, de todo modo, que trabalhos futuros investiguem o motivo dessa discrepância.

5.3 Pedágio controlado pela proporção υ de motoristas que são usuários

Nesta seção falamos sobre os resultados obtidos para experimentos variando o parâmetro υ para controlar a proporção de motoristas que são usuários. Demonstramos de que forma valores intermediários de υ influenciaram no desempenho do sistema.

A Tabela 4 mostra os valores $\overline{v}$ encontrados para cada rede nos casos em que

	υ		
Rede	0.25	0.5	0.75
B^1	17.8049 (0.059)	16.2575 (0.032)	15.3092 (0.014)
B^2	27.4718 (0.372)	25.8283 (0.052)	24.3809 (0.030)
B^3	36.9383 (0.761)	35.6004 (0.124)	33.9074 (0.038)
B^4	46.3341 (0.867)	45.1561 (0.345)	43.5885 (0.053)
B^5	55.6055 (0.819)	54.7135 (0.335)	53.2882 (0.090)
B^6	65.4056 (0.466)	64.4358 (0.257)	63.0141 (0.097)
B^7	75.1302 (0.297)	74.0701 (0.146)	72.7485 (0.091)
BB^1	8.9038 (0.040)	8.1258 (0.026)	7.6547 (0.012)
BB^3	21.4020 (0.104)	20.4526 (0.035)	19.6024 (0.025)
BB^5	49.3900 (0.043)	48.4149 (0.035)	47.6015 (0.027)
BB^7	122.9648 (0.071)	121.9543 (0.056)	121.1704 (0.068)
OW	66.9710 (0.003)	66.9692 (0.003)	66.9734 (0.005)
Anaheim	12.5906 (10^{-3})	12.5660 (10^{-3})	12.5534 (10^{-3})
EM	0.4334 (10^{-4})	0.4293 (10^{-4})	0.4275 (10^{-4})

Tabela 4: Valores $\overline{v}$ (com desvio padrão) para casos intermediários de υ.

$0 < \upsilon < 1$. O desvio padrão pode ser observado em parênteses.

Em quase todos os casos, os valores $\overline{v}$ intermediários estiveram entre $\overline{v}_{\text{UE}}$ e $\overline{v}_{\text{SO}}$. A única exceção a esse padrão foi a rede OW, cujos casos intermediários resultaram em valores $\overline{v}$ melhores do que o $\overline{v}_{\text{UE}}$. De todo modo, em nenhum momento um $\overline{v}$ para $\upsilon > 0$ (incluindo o $\overline{v}_{\text{SO}}$, quando $\upsilon = 1$) resultou em desempenho pior do que o de $\overline{v}_{\text{UE}}$ na mesma rede.

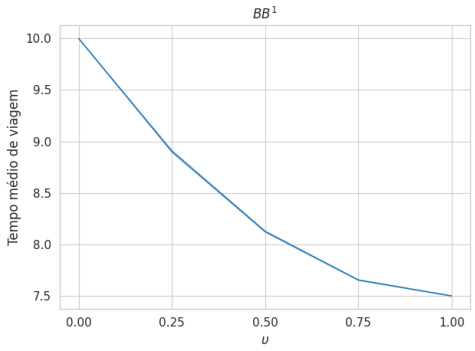

Figura 5: BB^1, exemplo de curva "côncava" nos experimentos de υ.

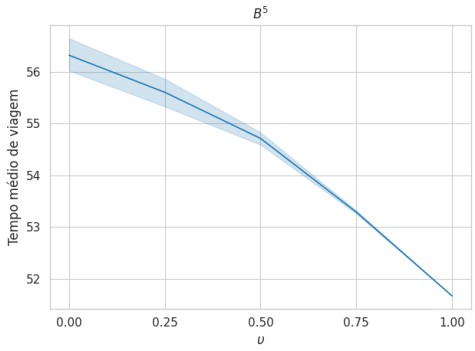

Figura 6: B^5, exemplo de curva "convexa" nos experimentos de υ.

Ao formatar os resultados em gráficos lineares quatro grupos emergiram com base no formato da linha, exemplificados por suas redes mais flagrantes nas Figuras 5, 6, 7 e 8. As áreas sombreadas em torno das linhas representam o intervalo de confiança de 95%.

A Figura 5 representa as redes com melhorias "côncavas" (constituídas por BB^1, BB^5, BB^7 e *Anaheim*), porque o $\Delta\overline{v}$ diminui à medida que υ se aproxima de 1, sendo

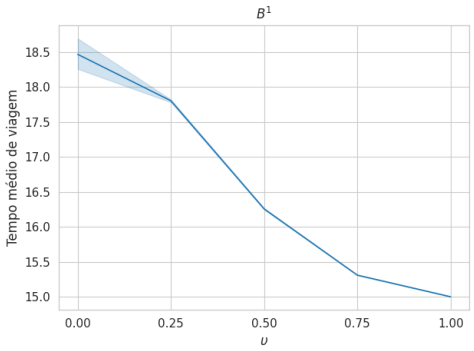

Figura 7: B^1, exemplo de curva sinuosa nos experimentos de υ.

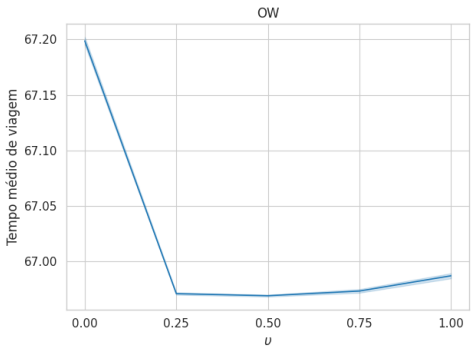

Figura 8: OW, o *outlier* dos experimentos de υ.

maior no começo do que no final. Isso significa dizer que, no geral, a diferença entre valores de $\overline{v}$ é maior em intervalos de υ próximos de 0 do que de 1. São, portanto, redes mais vulneráveis ao efeito de rendimentos decrescentes. A Figura 6 representa as redes com melhorias "convexas" (B^4, B^5, B^6 e B^7), cujo $\Delta\overline{v}$ aumenta à medida que υ se aproxima de 1. A Figura 7 representa as redes cujas melhorias realizam uma curva sinuosa (B^1, B^2, B^3, BB^3 e EM), em que $\Delta\overline{v}$ aumenta até $\upsilon = 0.5$ e depois passa a diminuir. As redes encontram-se distribuídas ao longo desses três grupos de maneira aproximadamente equilibrada, com apenas o grupo de curva sinuosa tendo uma rede a mais.

A rede OW, na Figura 8, é um *outlier*, com o melhor valor $\overline{v}$ ocorrendo em $\upsilon = 0.5$ ($\Delta\overline{v}$ positivo) e regredindo a partir daí ($\Delta\overline{v}$ negativo). Em nenhuma outra rede houve regressão do $\overline{v}$ em si, significando que nas demais os resultados foram sempre melhores à medida que o υ aumenta. Observamos que todos os valores $\overline{v}$ intermediários foram melhores do que o $\overline{v}_{\mathrm{SO}}$. Supondo ser improvável que essas proporções parciais de fato produzem um tempo de viagem médio melhor do que quando todos pagam pedágio, isso pode indicar que a simulação conseguiu convergir mais rápido nos casos intermediários do que ao realizar a aproximação de SO.

A ausência de valores $\overline{v}$ intermediários piores do que o $\overline{v}_{\mathrm{UE}}$ é evidência de que a aplicação parcial do sistema de pedágios não prejudicaria o fluxo de trânsito, contradizendo a ideia de que seria pior aplicar pedágios de forma parcial do que não aplicá-los. Além disso, observa-se melhorias de desempenho mesmo em proporções reduzidas, não sendo necessário atingir uma "massa crítica" de usuários para obter resultados. As redes que apresentaram curvas convexas e sinuosas (ver Figuras 6 e 7) foram as que mais chegaram perto de mostrar alguma estagnação inicial, porém sem grande significância.

Supondo uma adesão em função da quantidade de motoristas que são usuários, tudo indica que a adesão ao sistema de pedágios pode ser feita de maneira gradual. Não é necessário que ele seja implementado de uma só vez para obtermos resultados. Ele aparenta ser relativamente robusto à não-cooperação de pelo menos alguns agentes, e

em nenhum caso os efeitos da não-cooperação introduzem um comportamento caótico ao sistema. As melhorias são graduais e, no geral, consistentes.

5.4 Pedágio controlado pela proporção ρ de elos mais movimentados

Nesta seção falamos sobre os resultados obtidos em experimentos variando a proporção ρ de elos mais movimentados, com uma subseção para cada uma das duas formas de se controlar os pedágios em função de ρ. Na Subseção 5.4.1 apresentamos os resultados para o modo "rota" de pagamento, o primeiro a ser implementado. Apesar de ter obtido os melhores resultados dentre os dois modos, ele parte de suposições menos intuitivas. Na Subseção 5.4.2 apresentamos os resultados para o modo "elo" de pagamento, que foi implementado em partindo de premissas mais intuitivas. Porém, ele apresentou resultados significativamente piores do que o modo rota e também do que o controle por υ.

5.4.1 Modo de pagamento: rota

Em um primeiro instante determinou-se que o usuário pagaria pedágio para toda a rota mesmo que apenas um elo se encontre entre os mais movimentados. O motivo principal seria para fins de simplificação do algoritmo e também otimização, para que valores pudessem ser armazenados em caches e menos recálculo fosse necessário. Assim também não divergiríamos muito da conceitualização de Ramos et al. (2020a), que imaginou que na prática teríamos dispositivos individuais para cada motorista que calculassem o valor do pedágio apenas no final da viagem. Nesta subseção, apresentamos os resultados obtidos em experimentos partindo desses princípios.

A Tabela 5 lista os valores $\overline{v}$ obtidos para cada rede nos testes com valores intermediários de ρ em modo rota de pagamento. Os desvios padrão se encontram entre parênteses.

O primeiro ponto a ser observado é que em quase todas as redes os resultados de uma mesma rede foram muito similares entre si independente do valor de ρ, no caso de BB^1 sendo *exatamente* iguais. Dizemos que dois ou mais resultados são similares quando a diferença entre eles se encontra dentro do desvio padrão. Esse comportamento encontra-se exemplificado na Figura 9, em que há uma queda brusca de $\overline{v}_{\mathrm{UE}}$ para o primeiro $\overline{v}$ intermediário e segue-se adiante uma linha plana. Ou seja, um $\Delta\overline{v}$ grande no primeiro momento seguido de valores de $\Delta\overline{v}$ praticamente nulos. A única rede destoante foi a EM, cujo resultado para $\rho = 0.25$ foi levemente pior (fora do desvio padrão) do que para os demais valores de ρ, o que é visível na Figura 10. De todo modo, EM apresenta o mesmo comportamento de linha plana para os valores de ρ subsequentes.

	ρ, $m =$ ROTA		
Rede	0.25	0.5	0.75
B^1	15.0000 (10^{-6})	15.0000 (10^{-4})	15.0001 (10^{-4})
B^2	23.3334 (10^{-4})	23.3335 (10^{-4})	23.3339 (0.002)
B^3	32.5003 (10^{-3})	32.5003 (10^{-3})	32.5009 (0.002)
B^4	42.0005 (0.001)	42.0003 (10^{-3})	42.0007 (10^{-3})
B^5	51.6675 (0.001)	51.6673 (10^{-3})	51.6675 (10^{-3})
B^6	61.4308 (0.002)	61.4291 (10^{-3})	61.4306 (0.002)
B^7	71.2588 (0.007)	71.2516 (0.001)	71.2561 (0.006)
BB^1	7.5000 (0.000)	7.5000 (0.000)	7.5000 (0.000)
BB^3	19.0003 (10^{-3})	19.0006 (0.002)	19.0003 (10^{-3})
BB^5	47.0032 (0.002)	47.0034 (0.002)	47.0028 (0.002)
BB^7	120.5369 (0.066)	120.5908 (0.140)	120.5558 (0.073)
OW	66.9880 (0.007)	66.9869 (0.005)	66.9865 (0.005)
Anaheim	12.5474 (10^{-3})	12.5474 (10^{-3})	12.5476 (10^{-3})
EM	0.4275 (10^{-4})	0.4261 (10^{-4})	0.4261 (10^{-4})

Tabela 5: Valores $\overline{v}$ (com desvio padrão) para casos intermediários de ρ com $m =$ ROTA.

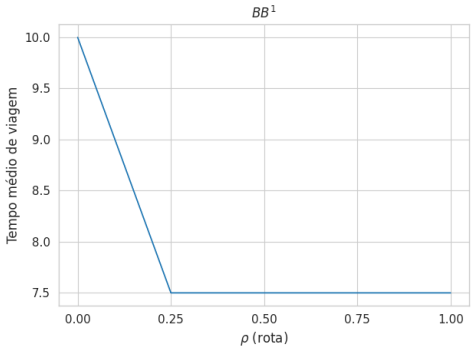

Figura 9: BB^1, exemplificando o padrão encontrado para quase todas as redes em experimentos com ρ em modo rota.

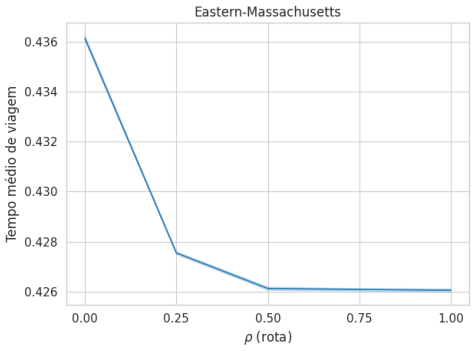

Figura 10: Eastern-Massachusetts, a única rede que para valores intermediários de ρ em modo rota divergiu significativamente (em $\rho = 0.25$) da aproximação de SO.

O segundo ponto (tornado mais visível pelos gráficos) é que os valores intermediários foram não só similares entre si, como também similares ao $\overline{v}_{\text{SO}}$ da rede. Novamente, as diferenças encontram-se dentro do desvio padrão. Isso nos leva à conclusão de que no modo elo de pagamento valores de ρ significativamente menores do que 1 já são o suficiente para o sistema apresentar o mesmo comportamento de quando todos os motoristas estão pagando pedágios. É bem possível que seja isso mesmo que esteja acontecendo, dando-se pela combinação de dois fatores.

Relembremos que no modo rota de pagamento basta que um elo da rota encontre-se entre os mais movimentados do sistema para que o pedágio seja cobrado por todo o trajeto. O primeiro fator é a possibilidade da rede ter uma proporção de elos que sejam

o suficiente para abarcar intersecções entre todas as rotas da rede. Se todos esses elos estiverem entre os ϱ elos mais movimentados, todas as rotas passarão por pelo menos um deles, portanto todas as rotas estarão sujeitas a pedágio pelo percurso todo, o que significa pedágio integral independente da escolha do motorista. O segundo fator é que os elos mais movimentados costumam ser justamente os que formam intersecções entre múltiplas rotas, já que eles concentrarão o fluxo de motoristas de múltiplas rotas. Os dois fatores em conjunto torna pedágios quase que inevitáveis nesse modo.

Considerando os fatores apresentados, podemos concluir que em todas as redes exceto a EM há uma proporção de elos menor ou igual do que 0.25 que constituem intersecções entre todas as rotas. No caso da rede EM, essa proporção provavelmente se encontra entre 0.25 e 0.5, o que significaria que a rede possui relativamente menos gargalos do que as demais.

Veremos adiante os resultados para quando os pedágios são aplicados apenas considerando os elos mais movimentados pelos quais o motorista passou.

5.4.2 Modo de pagamento: elo

Nesta subseção apresentamos os resultados dos experimentos para o modo elo. Considerando os resultados contra-intuitivos obtidos com o modo rota de pagamento, realizamos experimentos com outro modo de pagamento que poderia ser mais fácil de corresponder a situações reais – por exemplo, servindo para representar pontos de pedágio automático. Com o modo elo de pagamento o motorista paga pedágio correspondente *apenas* aos elos pelos quais ele passou que se encontram dentro do limiar, o que seria também mais justo.

Dessa forma o pedágio se torna menos penalizante, ao mesmo tempo em que fornece informações de maior resolução sobre a rota, permitindo que motoristas escolham com mais precisão. Por exemplo, rotas com menos elos dentre os mais movimentados tenderão a ter pedágios menores por conta dos demais elos onde o pedágio é zerado, isso permite que os motoristas indiretamente escolham entre rotas com mais ou menos elos com pedágio, o que também significa escolher mais elos menos movimentados.

A Tabela 6 lista os valores $\overline{v}$ obtidos para cada rede nos testes com valores intermediários de ρ em modo rota de pagamento. Os desvios padrão encontram-se entre parênteses. Os resultados do modo elo são melhores compreendidos quando comparados aos resultados de v e aos valores $\overline{v}_{\mathrm{UE}}$, portanto a Tabela 7 foi criada para facilitar essas comparações.

Os valores $\overline{v}$ na variação por ρ em modo elo de pagamento foram melhores do que na variação por v (para os mesmos valores de v e ρ) em apenas um terço dos casos, principalmente em redes sintéticas, mas nunca em rede abstraída de dados reais (Anaheim

	ρ, $m =$ ELO		
Rede	0.25	0.5	0.75
B^1	15.6751 (0.008)	18.5628 (0.577)	18.2412 (0.744)
B^2	26.3650 (0.560)	27.2738 (1.121)	24.7988 (0.393)
B^3	34.5061 (0.741)	37.9085 (1.177)	33.2568 (0.290)
B^4	44.7860 (0.349)	46.6335 (1.110)	42.3959 (0.191)
B^5	54.2742 (0.487)	56.1821 (0.547)	51.8754 (0.097)
B^6	64.3160 (0.363)	66.1502 (0.687)	61.5656 (0.058)
B^7	74.1701 (0.410)	75.9056 (0.410)	71.3602 (0.068)
BB^1	10.0000 (0.000)	10.0000 (0.000)	8.6667 (1.269)
BB^3	21.9725 (0.259)	21.7439 (0.581)	19.0007 (0.002)
BB^5	50.5410 (0.082)	50.5687 (0.038)	47.0026 (10^{-3})
BB^7	124.1161 (0.061)	124.1411 (0.149)	120.5444 (0.074)
OW	68.3217 (0.094)	67.2999 (0.005)	66.9871 (0.005)
Anaheim	13.2520 (0.008)	12.9589 (0.003)	12.6207 (0.002)
EM	0.4489 (0.001)	0.4447 (10^{-4})	0.4292 (10^{-3})

Tabela 6: Valores $\bar{v}$ (com desvio padrão) para casos intermediários de ρ com $m =$ ELO.

e EM) ou na rede OW (a mais complexa das redes sintéticas). Comparando em uma mesma rede os resultados de casos em que υ e ρ são iguais, nunca os resultados de ρ foram consistentemente melhores do que os de υ, porque em todas as redes o resultado de ρ é pior pelo menos nos casos em que $\rho, \upsilon = 0.5$. Os resultados no caso $\rho = 0.25$ se demonstraram melhores do que no caso $\upsilon = 0.25$ em todas as redes de classe B^p, mas apenas nelas. No caso $\rho = 0.75$ ele se demonstrou melhor do que $\upsilon = 0.75$ em 4 das 7 redes de classe B^p e em 3 das 4 redes de classe BB^p, concentrando-se em valores de p maiores.

Esse foi o único método que apresentou risco de sua implementação parcial causar maior congestionamento do que simplesmente não cobrar pedágios. Ou seja, dentre os múltiplos controles de pedágio considerados, o modo elo foi o único que apresentou valores $\bar{v}$ intermediários significativamente piores do que o valor $\bar{v}_{\text{UE}}$ da rede. As redes OW, Anaheim e EM obtiveram resultados piores do que o $\bar{v}_{\text{UE}}$ nos casos em que $\rho = 0.25$ e $\rho = 0.5$. Três redes sintéticas além da OW demonstraram no caso $\rho = 0.5$ desempenho pior do que a aproximação de UE. Em nenhuma rede o desempenho foi pior do que o UE de referência no caso $\rho = 0.75$; em contrapartida, como visto anteriormente, no mesmo valor 0.75 o controle por υ costuma obter resultados melhores.

Gerando gráficos lineares a partir dos resultados é possível observar três padrões gerais, exemplificados pelas Figuras 11, 13, 14. Comentemos cada padrão em sucessão.

A Figura 11 exemplifica o grupo de redes cujos resultados formam como que um "penhasco", porque o $\bar{v}$ para $\rho = 0.75$ difere enormemente dos resultados para outros casos intermediários de ρ. Formado pelas redes da classe BB^p, BB^1 (ver Figura 12) foi a

			Valor intermediário		
Rede		∼UE	0.25	0.5	0.75
B^1	ρ	18.4697	*15.6751*	**18.5628**	18.2412
	υ		17.8049	*16.2575*	*15.3092*
B^2	ρ	27.6516	*26.3650*	27.2738	24.7988
	υ		27.4718	*25.8283*	*24.3809*
B^3	ρ	37.3965	*34.5061*	**37.9085**	*33.2568*
	υ		36.9383	*35.6004*	33.9074
B^4	ρ	47.0528	*44.7860*	46.6335	*42.3959*
	υ		46.3341	*45.1561*	43.5885
B^5	ρ	56.3216	*54.2742*	56.1821	*51.8754*
	υ		55.6055	*54.7135*	53.2882
B^6	ρ	66.2312	*64.3160*	66.1502	*61.5656*
	υ		65.4056	*64.4358*	63.0141
B^7	ρ	75.9613	*74.1701*	75.9056	*71.3602*
	υ		75.1302	*74.0701*	72.7485
BB^1	ρ	10.0000	10.0000	10.0000	8.6667
	υ		*8.9038*	*8.1258*	*7.6547*
BB^3	ρ	22.0091	21.9725	21.7439	*19.0007*
	υ		*21.4020*	*20.4526*	19.6024
BB^5	ρ	50.5608	50.5410	**50.5687**	*47.0026*
	υ		*49.3900*	*48.4149*	47.6015
BB^7	ρ	124.1569	124.1161	124.1411	*120.5444*
	υ		*122.9648*	*121.9543*	121.1704
OW	ρ	67.1986	**68.3217**	**67.2999**	66.9871
	υ		*66.9710*	*66.9692*	*66.9734*
Anaheim	ρ	12.6300	**13.2520**	**12.9589**	12.6207
	υ		*12.5906*	*12.5660*	*12.5534*
EM	ρ	0.4362	**0.4489**	**0.4447**	0.4292
	υ		*0.4334*	*0.4293*	*0.4275*

Tabela 7: Valores $\overline{v}$ intermediários de ρ em modo elo de pagamento, comparados em cada rede com seu valor $\overline{v}_{\text{UE}}$ e valores $\overline{v}$ de υ. Valores $\overline{v}$ piores do que $\overline{v}_{\text{UE}}$ em **negrito**. Melhor resultado entre casos de uma mesma rede com υ e ρ iguais em *itálico* e fundo colorido.

única do grupo que no caso $\rho = 0.75$ obteve resultado pior do que para $\upsilon = 0.75$.

A Figura 13 exemplifica o grupo de redes cujos resultados formam um desenho que lembra um "relâmpago", cujos resultados para valores intermediários alternam entre melhora e piora relativo ao resultado anterior. É formado pelas redes da classe B^p.

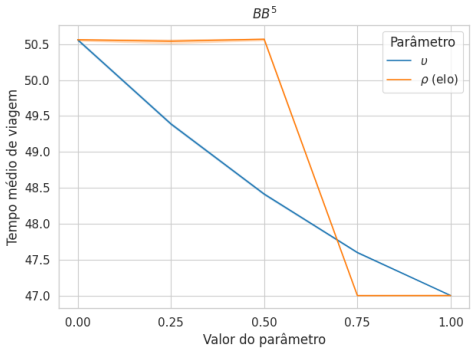

Figura 11: BB^5, exemplo de resultados com figura de "penhasco" nos experimentos de ρ, $m =$ ELO.

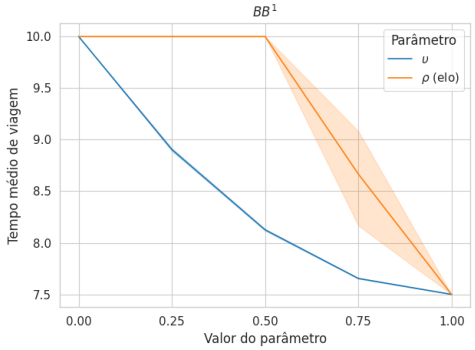

Figura 12: BB^1, única rede com padrão "penhasco" dos experimentos de ρ, $m =$ ELO que obteve resultado para $\rho = 0.75$ pior do que para $\upsilon = 0.75$.

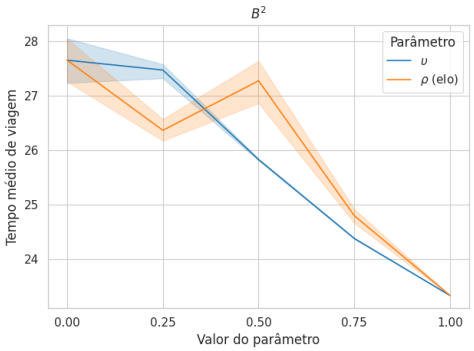

Figura 13: B^2, exemplo de resultados com figura de "relâmpago" nos experimentos de ρ, $m =$ ELO.

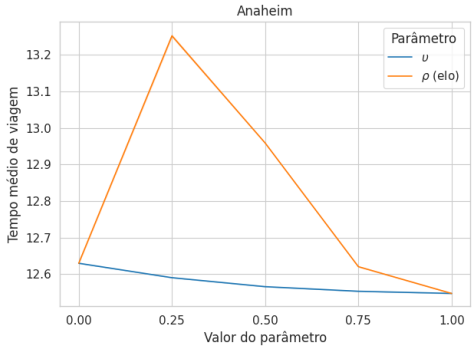

Figura 14: Anaheim, exemplo de resultados com figura de "montanha" nos experimentos de ρ, $m =$ ELO.

Por fim, a Figura 14 exemplifica o grupo de redes cujos resultados formam como que uma "montanha", porque o $\overline{v}$ obtido para $\rho = 0.25$ eleva-se significativamente acima da aproximação de UE. Ele é constituído pelas redes OW, Anaheim e EM, justamente as únicas redes com resultados piores do que $\overline{v}_{\text{UE}}$ no caso $\rho = 0.25$.

Há algumas possíveis explicações para o comportamento observado do modo elo de pagamento. Talvez ele exija mais episódios para convergir, dado que o método de controle de pedágios introduz significativamente mais dinamismo à rede. Afinal, a cada episódio os elos mais movimentados são recalculados e, diferente do modo rota, as escolhas dos motoristas são muito mais impactadas pela atualização.

Porém, isso também introduz o risco de que a variação de elos mais movimentados acabe se dando de maneira cíclica. Se desconsiderarmos os fatores de decaimento o sistema nunca se estabilizaria. Por exemplo: motoristas podem aprender a evitar certos elos, mas então novos elos se tornam os mais movimentados, e para evitar esses novos mais movimentados eles voltam a frequentar os antigos, que voltam a ser os mais movimentados que eles aprenderam a evitar inicialmente, e assim sucessivamente. Considerando

os fatores de decaimento, o sistema acaba se estabilizando apenas porque os motoristas se tornam incapazes de aprender a partir de certo ponto, mas sem terem chegado mais perto do ótimo. De jeito ou de outro, o aprendizado se torna muito mais difícil para os motoristas.

Para diminuir a variabilidade da simulação e prevenir-se contra ciclos a heurística de determinação dos elos mais movimentados poderia ser alterada. Em uma situação real, a cidade que está adotando um sistema de pedágios gradualmente dificilmente teria os recursos para determinar as ruas mais movimentadas todos os dias.

O mais realista de se acontecer, como no caso de Leape (2006), é o seguinte desenvolvimento: 1) o sistema já encontra-se relativamente estabilizado próximo do UE; 2) pontos para pedágio levando em conta as concentrações de fluxo são determinados a partir desse estado em equilíbrio; 3) novas determinações não são feitas enquanto o sistema não se estabilizar novamente. Quando novas determinações forem feitas pode-se muito bem escolher entre manter os pontos de pedágio que já existem e expandir, ou então trocar as localizações de pontos de pedágio para não aumentar a quantidade de pontos que obrigam o pagamento de pedágios. A implementação de tais experimentos em ambientes simulados ficaria para trabalhos futuros.

Mas os fatores comentados não são os únicos a serem levados em conta. Estabelecer pedágio apenas nos elos mais movimentados induz os motoristas a simplesmente evitarem aumentar o fluxo de onde o fluxo já é relativamente grande, mas ele não será penalizado de forma alguma por aumentar o fluxo em outros elos. Isso pode fazer com que o congestionamento diminua apenas numa área localizada enquanto que ele piora em todo o restante da rede. Leape (2006) verificou que não houve muita mudança no trânsito de Londres em geral e pioras na área em torno do centro apenas no primeiro ano de implementação, depois ficando melhor do que estava antes. De todo modo, o risco dessa piora ainda pode ser melhor investigado.

6 Conclusão

Neste trabalho apresentamos os resultados de experimentos com uma nova versão do algoritmo TQ-*learning*, cujo diferencial é permitir controlar o pagamento de pedágios com base em condições circunstanciais. Nosso objetivo foi avaliar o desempenho do algoritmo em diferentes níveis de adesão ao sistema de pedágios e suas implicações para a implementação prática, proporcionando uma análise mais aprofundada do custo-benefício do TQ-*learning*. Com a flexibilidade das condições implementadas, testamos e comparamos três formas de controle e pagamento de pedágios, constituindo diferentes cenários de adesão gradual.

As principais questões a serem investigadas eram: 1) a viabilidade da implementação de um TQ-*learning* com pagamento circunstancial; 2) os riscos da adesão parcial do pagamento de pedágios trazer mais prejuízos do que nenhuma adesão, em diferentes cenários; 3) a necessidade de adesão total aos pedágios para obter resultados satisfatórios, em diferentes cenários; e 4) qual forma de controle e pagamento de pedágios teria maior valor na formulação de políticas públicas.

No primeiro cenário o pedágio é controlado por υ, a proporção de motoristas que são usuários e sempre pagam pedágios. No segundo cenário o pedágio é controlado por ρ, a proporção de elos mais movimentados que tornam o pedágio obrigatório ao serem visitados em trajeto, com o modo "rota" de pagamento, significando que ao se passar por um elo dentro do limiar de ρ o pedágio é calculado pro trajeto inteiro. No terceiro cenário o pedágio também é controlado por ρ, mas usa-se o modo "elo" de pagamento, e o pedágio é calculado apenas para as seções do trajeto que constituem os elos dentro do limiar de mais movimentados.

O controle por υ é o mais parcimonioso em suas premissas e foi o que trouxe resultados mais estáveis e de mais fácil interpretação. Ele demonstrou os menores riscos, pois em nenhum momento seus valores intermediários levaram a resultados piores do que a aproximação de UE para uma dada rede. Com exceção da rede OW, os resultados pra valores intermediários maiores não são piores do que pra valores intermediários menores. Ou seja, as melhorias são, no geral, graduais e consistentes, sem grande risco de piora durante a progressão.

O ponto fraco do controle por υ é que, por ser parcimonioso, é também muito abstrato, e não necessariamente modelaria a progressão da adesão a um sistema de pedágios real. Todavia, ele indica que há vantagem em um sistema de pedágio real controlar (se possível) sua adesão pelos motoristas ao invés de por trajetos específicos. Um exemplo de cenário real seria a cobrança de pedágios por dispositivos instalados nos carros dos

próprios motoristas, como sugerido originalmente por Ramos et al. (2020a). As simulações com υ poderiam ajudar a determinar a robustez do sistema para evasões. De fato, os resultados já obtidos indicam que evasões por si só não representariam riscos de pioras descontroladas, mesmo que, claro, diminuam o desempenho do sistema.

Em um trabalho futuro, seria possível alterar o comportamento dos agentes para permitir que eles mesmos decidam, com base em critérios pré-estabelecidos, se serão usuários ou não. Dessa forma, seria possível investigar e possivelmente descobrir as condições que levariam os motoristas a uma adesão voluntária ao sistema de pedágios.

Foi para tentar lidar com casos mais concretos que as abordagens de ρ foram implementadas. O controle por ρ em modo rota de pagamento teve o melhor desempenho dos cenários, mas a custo de ser o menos parcimonioso dos cenários e o mais difícil de ser correspondido na realidade. Dado que na prática ele não foi diferente de uma aproximação de SO básica, seu comportamento foi o menos interessante. Ainda assim, a razão desse comportamento nos forneceu informações valiosas sobre as redes de tráfego testadas. Encontramos nas redes a tendência de que a interseção de todas as rotas formem uma proporção relativamente pequena de elos (na maioria das vezes, menor do que 25%), o que talvez seja comum em malhas reais no geral. Dado que todas as rotas passam por esse conjunto, o tráfego tende a se concentrar justamente nesses elos.

O último cenário a ser investigado foi o controle por ρ em modo elo de pagamento, que obteve os piores e mais imprevisíveis desempenhos. Foi o único que obteve resultados piores do que a aproximação de UE, inclusive significativamente piores. Ele indica que há de fato um risco na adesão parcial de pedágios, a depender da forma como a adesão é condicionada.

As abordagens de pedágios baseadas em ρ (independente do modo de pagamento) introduzem um possível elemento de centralização, caso seja necessário que as observações de fluxo de tráfego sejam enviados a alguma central para determinar quais os elos mais movimentados. Subsequentemente, seria necessário acessar essa central para obter tal informação.

O trabalho de Ramos et al. (2020a) tem como objetivo expresso desenvolver um sistema que seja o mais descentralizado possível. Para ater-se a esse ideal, na prática seria recomendado obter alguma maneira descentralizada de circular as informações. Em um cenário real é possível imaginar uma prefeitura que, tendo acesso a esses dados, ao invés de usar dispositivos instalados nos carros se valeria de sensores de pagamento automático distribuídos em pontos estratégicos da cidade. De todo modo, é necessário maior investigação para demonstrar que a abordagem dos elos mais movimentados é de fato viável, já que em nossos experimentos ou as premissas são difíceis de se corresponder na realidade (modo rota de pagamento) ou simplesmente não resultaram em bom desempenho (modo elo).

Apesar do controle por ρ ter sido uma tentativa de se aproximar de casos mais concretos, ele ainda assim corresponde pouco a cenários reais e trouxe resultados mais inconclusivos do que os do controle por v. Concluímos que, por enquanto, o controle da adesão de pedágios pela proporção de motoristas que são usuários é, dentre as experimentadas, a abordagem de maior valor para a formulação de políticas públicas.

Todavia, ainda vale a pena investigar mais a fundo abordagens mais concretas, o que exigiria maior estudo de técnicas já implementadas na vida real. Aqui se enquadra a sugestão para trabalhos futuros mencionada no final da Subseção 5.4.2, comentando sobre o caso de Leape (2006). No geral, pode ser útil realizar experimentos baseados em um ciclo sucessivo de alteração e estabilização. Isso poderia ser usado para investigar a progressão de valores intermediários de parâmetros em um mesmo experimento ao invés de em múltiplos experimentos para cada valor, seguindo um ritmo que dê ao sistema tempo o suficiente para se estabilizar antes de progredir no parâmetro.

Trabalhos futuros partindo de abordagens similares ao controle por ρ fariam bem em manter um registro das proporções de motoristas que pagaram pedágio em um dado episódio ou não, para assim determinar mais precisamente a eficácia do pagamento dos pedágios.

Neste trabalho não investigamos a interação entre os diversos cenários, mas isso pode ser útil para expandir as opções ao se implementar um sistema de pedágios de forma gradual. Administradores não precisarem se limitar a apenas uma ou outra abordagem. A investigação de interações pode até mesmo ajudar a descobrir métodos que se complementam e são mais eficazes em conjunto do que isoladamente.

É interessante que a rede OW, apesar de ser uma rede sintética, apresentou características em comum com as redes de dados reais que não foram demonstradas pelas demais redes sintéticas. Observamos que ela: 1) aparenta também exigir uma maior quantidade de episódios para convergência; 2) requer mais tempo de processamento do que as demais redes sintéticas; e 3) seguiu o mesmo padrão de resultados que as redes Anaheim e Eastern-Massachusetts nos experimentos controlando por ρ em modo elo. É bem possível que mais semelhanças ainda possam ser encontradas. Isso indica que, por apresentar mais desafios, ela é uma rede sintética apropriada para experimentos mais rigorosos antes de partir para as redes de dados reais. As redes de classes B^p e BB^p, por serem variadas e de bem rápido processamento, ainda são úteis para provas de conceitos.

Tivemos a ideia de investigar como a consideração do consumo de combustível pelos motoristas poderia influenciar no congestionamento dos sistemas, dado que o TQ-*learning* trabalha apenas com tempo de viagem e pedágios. Como ele seria fatorado nas tomadas de decisões de agentes? Será que há alguma correspondência entre o consumo de combustível e o MCT? Ele poderia de alguma forma substituir ou complementar os pedágios no controle de congestionamentos? Porém, tais experimentos não seriam possíveis

com a mesma base de dados utilizada neste trabalho, pois, faltando a informação da distância entre entre vértices, não há como derivar o consumo de combustível apenas das demais informações que a base provê[1]. Deixamos essa como nossa última sugestão para trabalhos futuros, e o trabalho de Dias et al. (2014) poderia servir como referência.

Em suma, é possível criar novas versões de TQ-*learning* com pedágio circunstancial, exemplificado pelo nosso modelo. Apresentamos não apenas uma forma de controle da condição de pagamento, mas três. A mais parcimoniosa e consistente funcionou controlando uma proporção de motoristas que sempre pagam ou não os pedágios. O controle pelos elos mais movimentados trouxe resultados inconclusivos, pois ainda falta encontrar um equilíbrio entre desempenho e aplicabilidade na vida real, trazendo resultados muito divergentes a depender do modo de pagamento utilizado. Mais investigações em cima do controle por elos podem ser feitas, mas, por enquanto, o controle por usuários aparenta já ser bem eficaz e informativo.

Concluímos aqui a nossa dissertação. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD 22-23.

[1] Ver (Shang et al., 2014, p. 5).

Referências

Agogino, A. K. and K. Tumer

2004. Unifying temporal and structural credit assignment problems. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '04, P. 980–987, USA. IEEE Computer Society. Citado na página 39.

Beckmann, M., C. B. McGuire, C. B. Winsten, and T. C. Koopmans

1957. Studies in the economics of transportation. *The Economic Journal*, 67(265):116–118. Citado 2 vezes nas páginas 29 e 34.

Braess, D.

1968. Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung Operations Research - Recherche Opérationnelle*, 12:258–268. Citado na página 27.

Bureau of Public Roads

1964. *Traffic Assignment Manual*. Washington D.C.: U.S. Department of Commerce, Urban Planning Division. Citado na página 29.

Colby, M., T. Duchow-Pressley, J. J. Chung, and K. Tumer

2016. Local approximation of difference evaluation functions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, P. 521–529, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems. Citado na página 39.

D'Acierno, L., B. Montella, and F. De Lucia

2006. A stochastic traffic assignment algorithm based on ant colony optimisation. In *Ant Colony Optimization and Swarm Intelligence*, M. Dorigo, L. M. Gambardella, M. Birattari, A. Martinoli, R. Poli, and T. Stützle, eds., Pp. 25–36, Berlin, Heidelberg. Springer Berlin Heidelberg. Citado na página 37.

Dias, J. C., P. Machado, D. C. Silva, and P. H. Abreu

2014. An inverted ant colony optimization approach to traffic. *Engineering Applications of Artificial Intelligence*, 36:122–133. Citado 3 vezes nas páginas 37, 38 e 69.

Dorigo, M., A. Colorni, and V. Maniezzo

1991. Distributed optimization by ant colonies. Citado na página 37.

Francesco, M. D. and M. D. Rosini

2015. Rigorous derivation of nonlinear scalar conservation laws from follow-the-leader

type models via many particle limit. *Archive for Rational Mechanics and Analysis*, 217(3):831–871. Citado na página 29.

Hearn, D. W. and M. V. Ramana
1998. *Solving Congestion Toll Pricing Models*, Pp. 109–124. Boston, MA: Springer US. Citado 3 vezes nas páginas 23, 38 e 39.

Jabbarpour, M. R., H. Malakooti, R. M. Noor, N. B. Anuar, and N. Khamis
2014. Ant colony optimisation for vehicle traffic systems: Applications and challenges. *International Journal of Bio-Inspired Computation*, 6:32–56. Citado na página 38.

Joshi, D. J., I. Kale, S. Gandewar, O. Korate, D. Patwari, and S. Patil
1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1311 AISC:297–308. Citado na página 29.

Klügl, F., A. Bazzan, M. Lujak, I. Dusparic, F. Klügl, and G. Vizzari
2021. Accelerating route choice learning with experience sharing in a commuting scenario: An agent-based approach. *AI Commun.*, 34(1):105–119. Citado na página 41.

Knuth, D. E.
1992. Two notes on notation. Citado na página 27.

Koutsoupias, E. and C. Papadimitriou
1999. Worst-case equilibria. In *STACS 99*, C. Meinel and S. Tison, eds., Pp. 404–413, Berlin, Heidelberg. Springer. Citado na página 28.

Leape, J.
2006. The london congestion charge. *Journal of Economic Perspectives*, 20:157–176. Citado 3 vezes nas páginas 23, 65 e 68.

Legge, D.
2005. The strategic control of an ant-based routing system using neural net q-learning agents. In *Adaptive Agents and Multi-Agent Systems II*, D. Kudenko, D. Kazakov, and E. Alonso, eds., Pp. 147–166, Berlin, Heidelberg. Springer Berlin Heidelberg. Citado na página 41.

Lin, H., T. Roughgarden, E. Tardos, and A. Walkover
2011. Stronger bounds on braess's paradox and the maximum latency of selfish routing. *SIAM Journal on Discrete Mathematics*, 25(4):1667–1686. Citado 2 vezes nas páginas 11 e 51.

Mirzaei, H., G. Sharon, S. Boyles, T. Givargis, and P. Stone
2018. Enhanced delta-tolling: Traffic optimization via policy gradient reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Pp. 47–52. Citado na página 39.

Myerson, R.
1997. *Game Theory: Analysis of Conflict*. Harvard University Press. Citado na página 27.

Nash, J.
1951. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295. Citado na página 27.

Ortúzar, J. d. D. and L. Willumsen
2011. *Modelling Transport, Fourth Edition*. Citado na página 50.

Pigou, A. C.
1920. *The Economics of Welfare*. Routledge. Citado 2 vezes nas páginas 23 e 29.

Ramos, G. de. O.
2018. *Regret Minimisation and System-Efficiency in Route Choice*. PhD thesis, Universidade Federal do Rio Grande do Sul, Brazil. Citado 4 vezes nas páginas 52, 53, 54 e 76.

Ramos, G. de. O. and A. L. C. Bazzan
2015. Towards the user equilibrium in traffic assignment using grasp with path relinking. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, GECCO '15, P. 473–480, New York, NY, USA. Association for Computing Machinery. Citado 2 vezes nas páginas 11 e 51.

Ramos, G. de. O., B. C. da Silva, R. Rădulescu, A. L. C. Bazzan, and A. Nowé
2020a. Toll-based reinforcement learning for efficient equilibria in route choice. *Knowledge Engineering Review*. Citado 11 vezes nas páginas 24, 31, 32, 34, 35, 40, 50, 53, 56, 59 e 67.

Ramos, G. de. O., R. Rădulescu, A. Nowé, and A. R. Tavares
2020b. Toll-based learning for minimising congestion under heterogeneous preferences. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*, B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, and G. Sukthankar, eds., Pp. 1098–1106, Auckland, New Zealand. IFAAMAS. Citado 2 vezes nas páginas 33 e 41.

Shang, J., Y. Zheng, W. Tong, E. Chang, and Y. Yu
2014. Inferring gas consumption and pollution emission of vehicles throughout a city. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, P. 1027–1036, New York, NY, USA. Association for Computing Machinery. Citado na página 69.

Sharon, G., S. D. Boyles, S. Alkoby, and P. Stone
2019. Marginal cost pricing with a fixed error factor in traffic networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, P. 1539–1546, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems. Citado 6 vezes nas páginas 11, 51, 52, 54, 56 e 76.

Sharon, G., J. P. Hanna, T. Rambha, M. W. Levin, M. Albert, S. D. Boyles, and P. Stone
2017. Real-time adaptive tolling scheme for optimized social welfare in traffic networks. AAMAS '17, P. 828–836, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems. Citado na página 39.

Somuyiwa, A. O., S. O. Fadare, and B. B. Ayantoyinbo
2015. Analysis of the cost of traffic congestion on worker's productivity in a mega city of a developing economy. *International Review of Management and Business Research*, 4(3):644. Citado na página 23.

Stefanello, F., L. S. Buriol, M. J. Hirsch, P. M. Pardalos, T. Querido, M. G. C. Resende, and M. Ritt
2017. On the minimization of traffic congestion in road networks with tolls. *Annals of Operations Research*, 249(1):119–139. Citado 2 vezes nas páginas 39 e 56.

Sutton, R. S. and A. G. Barto
1998. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054. Citado 2 vezes nas páginas 29 e 31.

Wardrop, J.
1952. *Some Theoretical Aspects of Road Traffic Research*, Road paper. Institution of Civil Engineers. Citado na página 27.

Watkins, C. J. C. H.
1989. *Learning from Delayed Rewards*. Cambridge University. Citado na página 30.

Watkins, C. J. C. H. and P. Dayan
1992. Q-learning. *Machine Learning 1992 8:3*, 8:279–292. Citado na página 30.

Wolpert, D. H. and K. Tumer
1999. An introduction to collective intelligence. *CoRR*, cs.LG/9908014. Citado na página 39.

Wood, D. A., Standing Advisory Committee on Trunk Road Assessment, and Great Britain: Department of Transport
1994. *Trunk Roads and the Generation of Traffic*. H.M. Stationery Office. Citado na página 23.

Yen, J. Y.

1971. Finding the k shortest loopless paths in a network. *Management Science*, 17(11):712–716. Citado na página 33.

Zhong, N., J. Cao, and Y. Wang

2017. Traffic congestion, ambient air pollution, and health: Evidence from driving restrictions in beijing. *Journal of the Association of Environmental and Resource Economists*, 4(3):821–856. Citado na página 23.

Anexos

ANEXO A – Gráficos de resultados

Aqui apresentamos gráficos lineares diagramando os resultados de todos os experimentos para cada rede. Os valores ideais de UE e SO de cada rede, de acordo com a literatura (Ramos, 2018; Sharon et al., 2019), estão indicados por uma linha pontilhada vermelhas (UE) e uma preta (SO).

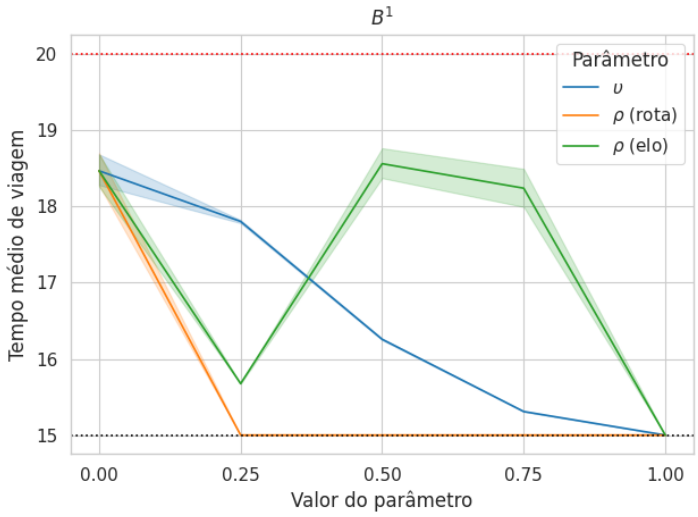

Figura 15: B^1

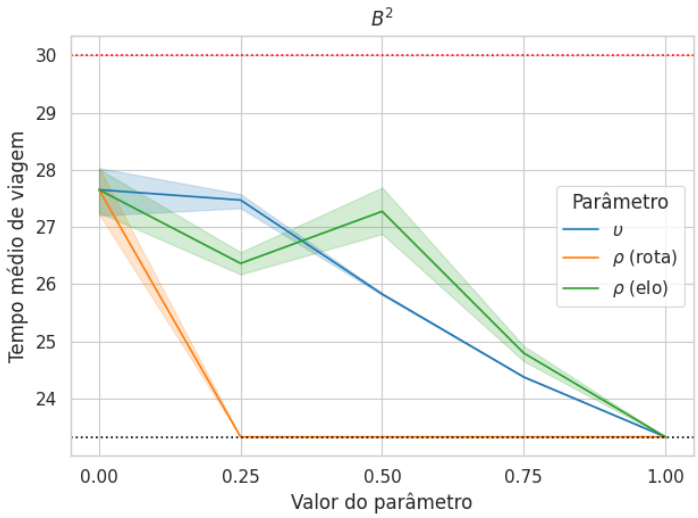

Figura 16: B^2

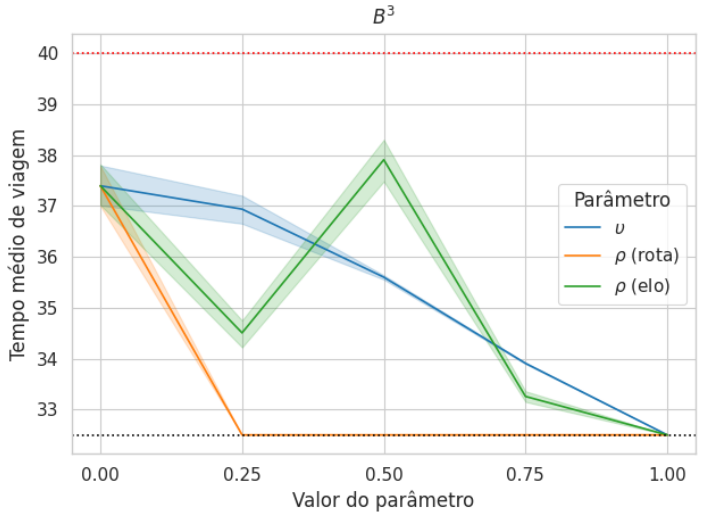

Figura 17: B^3

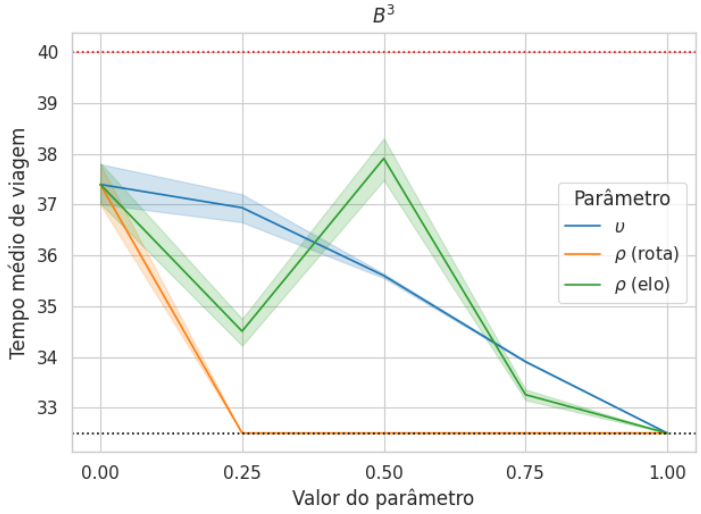

Figura 18: B^3

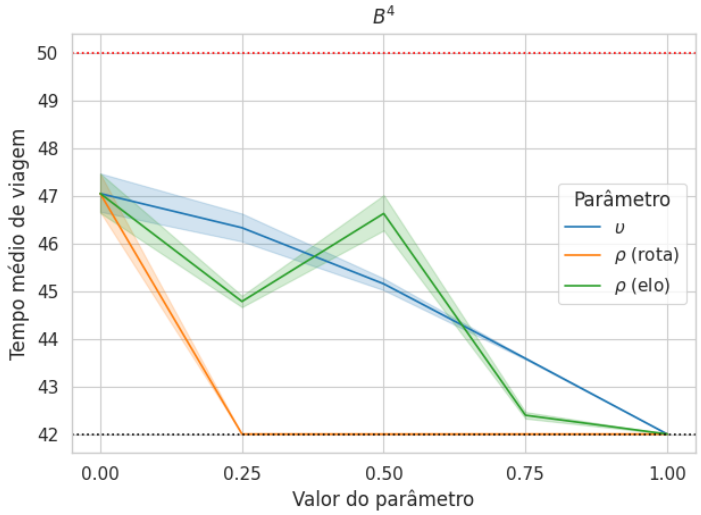

Figura 19: B^4

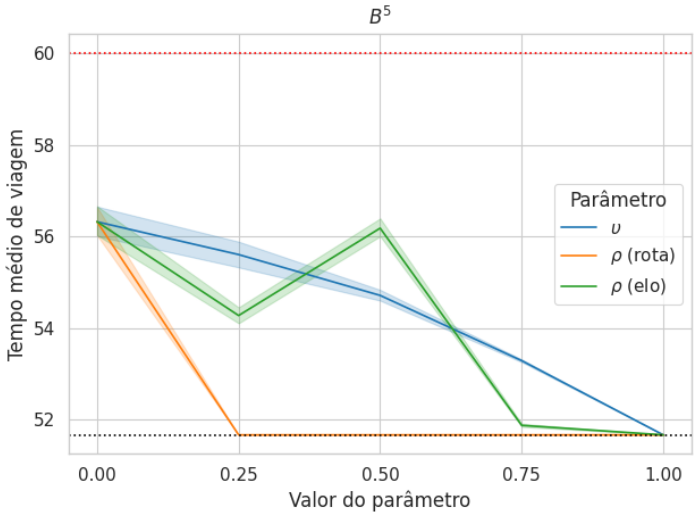

Figura 20: B^5

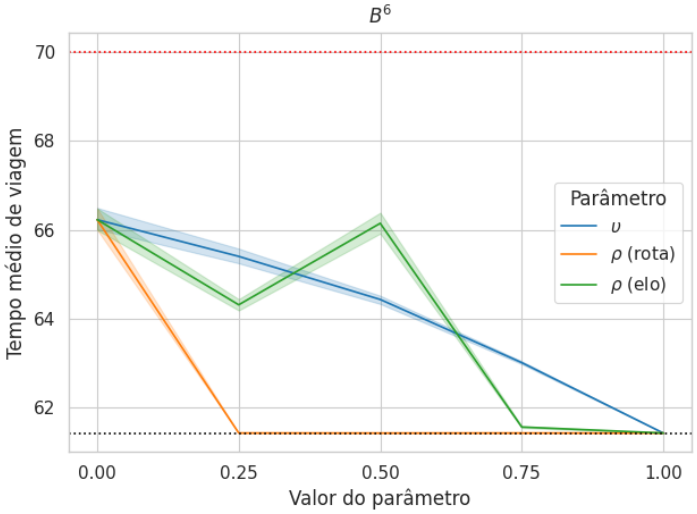

Figura 21: B^6

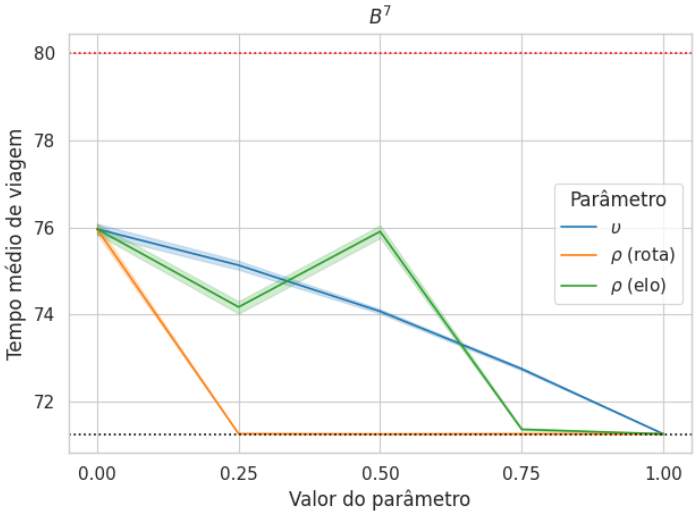

Figura 22: B^7

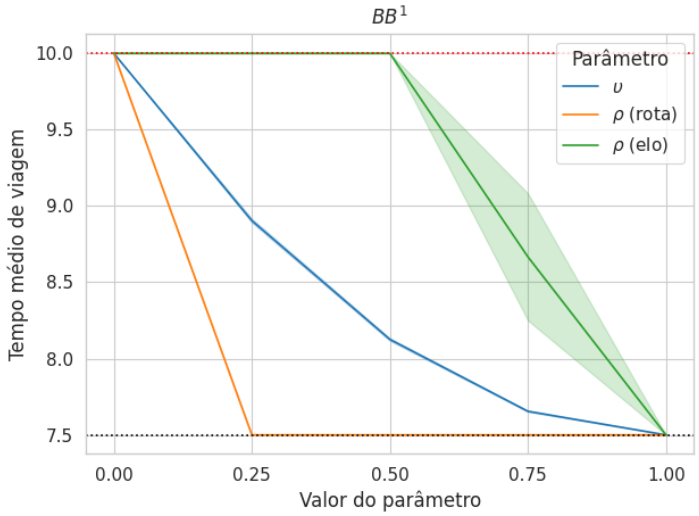

Figura 23: BB^1

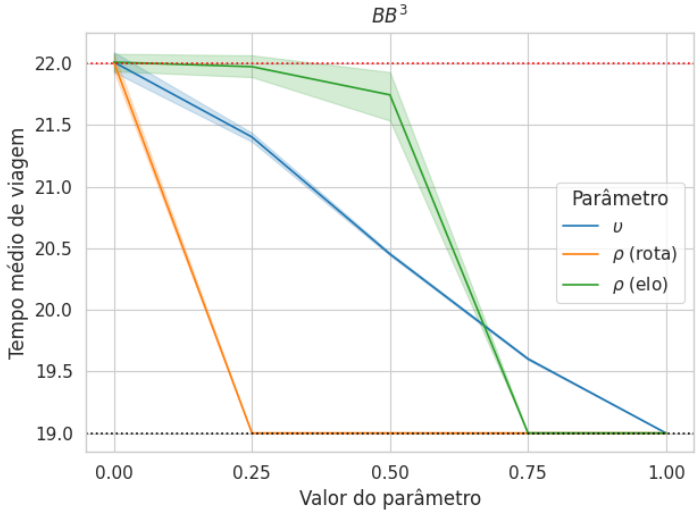

Figura 24: BB^3

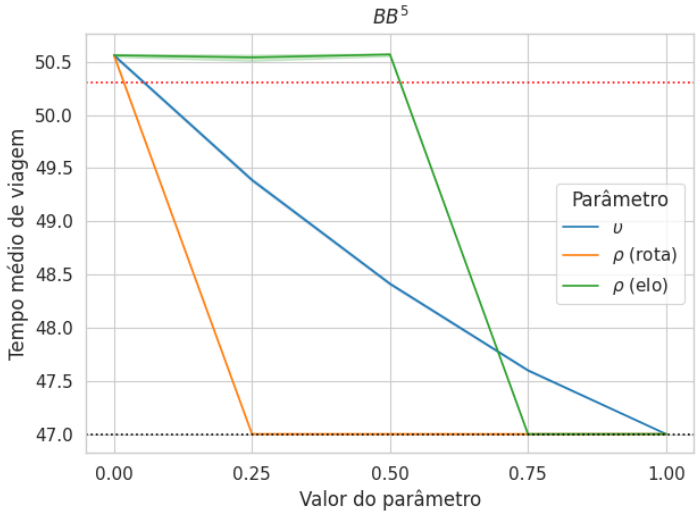

Figura 25: BB^5

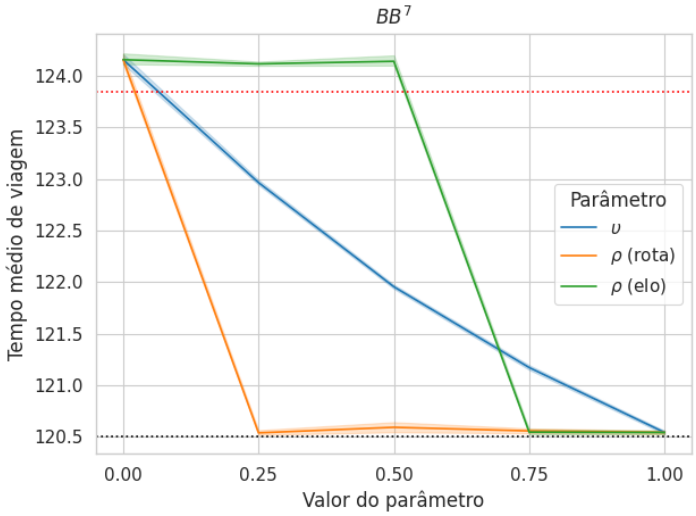

Figura 26: BB^7

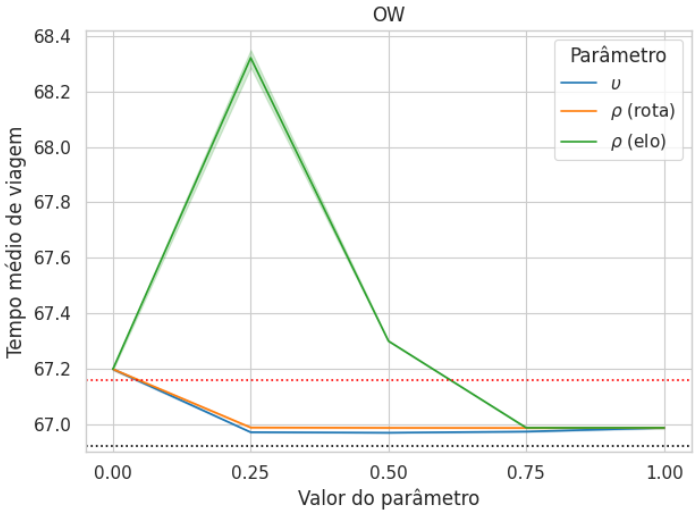

Figura 27: OW

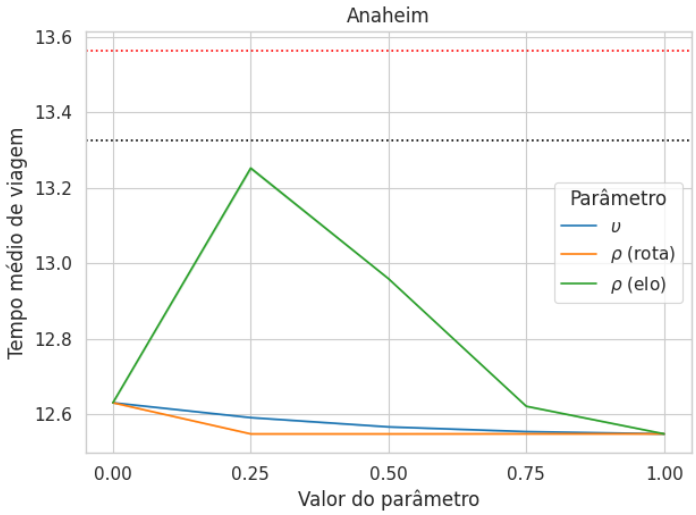

Figura 28: Anaheim

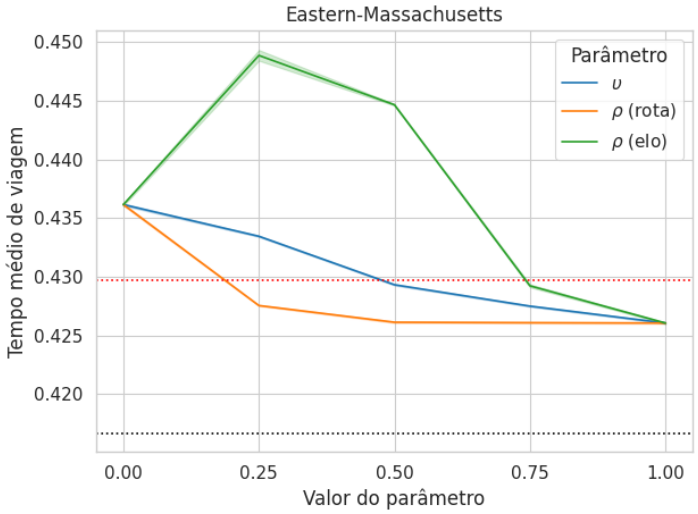

Figura 29: Eastern-Massachusetts