XVIII ANPET
XVIII CONGRESSO DE PESQUISA E ENSINO EM TRANSPORTES

AGENTES DE CONTROLE DISTRIBUÍDOS: UM ESTUDO DE CASO EM REDES DE TRÁFEGO VEICULAR URBANO

Eduardo Camponogara
Maurício R. G. Serra
Departamento de Automação e Sistemas
Universidade Federal de Santa Catarina

RESUMO
Este artigo examina a tecnologia de controle de tráfego veicular ora empregada e oferece uma nova perspectiva das políticas de operação, reformulando o problema dinâmico de operação como um jogo estocástico entre agentes de controle distribuídos. São discutidos aspectos relativos à otimalidade de políticas de controle e é proposto um algoritmo livre de modelo (*model-free*) para sintetizar políticas de controle, o qual foi projetado por meio de uma técnica de aprendizagem por reforço. Evidências numéricas coletadas a partir de simulações computacionais em uma rede representativa demonstram o potencial da tecnologia de aprendizagem de máquina para melhorar o desempenho das redes de tráfego veicular.

ABSTRACT
This paper surveys the existing traffic control technology, and offers a new perspective into the operation of traffic networks by formulating the operation problem as a distributed, stochastic game among distributed control agents. Issues regarding the optimality of control policies are discussed and a model-free, reinforcement-learning algorithm is proposed to search for optimal control policies. Numerical evidence, gathered from computer simulation of a representative traffic network, demonstrate the potential of machine-learning technology to improve the performance of traffic networks.

1. INTRODUÇÃO

Não raramente, motoristas são forçados a aguardar por períodos excessivamente longos nas interseções semaforizadas; usuários do transporte coletivo suportam longas jornadas nos sistemas de transporte públicos que são mal projetados ou operados de maneira ineficaz; e os direitos dos pedestres não são atendidos. Duas maneiras de se melhorar o desempenho dos sistemas de tráfego são: (i) expandir a infra-estrutura de tráfego, o que implica a construção de novos meios de transporte e a ampliação dos já existentes; e (ii) a implementação de tecnologias modernas de controle e de gerenciamento de tráfego que venham a elevar o desempenho das redes, o qual pode ser medido em termos da redução dos tempos de viagem, do aumento do bem-estar dos usuários e da redução nas emissões de dióxido de carbono. Ambas as alternativas apresentam vantagens e desvantagens que devem ser ponderadas e analisadas caso a caso: a primeira alternativa requer gastos significativos e pode não ser viável devido a limitações físicas ou ambientais; a segunda alternativa, por outro lado, não trará benefícios satisfatórios em um sistema saturado. Assim, a melhor linha de ação é uma combinação da expansão da infra-estrutura e da modernização da tecnologia, mas certamente a primeira alternativa não pode ser eficiente sem a última.

Apesar dos sucessos das pesquisas de campo durante os anos 80 e 90, as quais mostraram que ganhos significativos em desempenho podem ser produzidos com a implementação de tecnologias de controle em tempo real (Hunt et al., 1981; Thorpe e Anderson, 1996), a vasta maioria dos semáforos ainda são operados com base em planos de tempo fixo, conhecidos como controle pela hora do dia (Crabtree, Vincent e Harrison, 1996) e abreviados por TOD (*time of the day traffic control*). Os esforços para melhorar o controle TOD se concentraram em duas

frentes: (i) o controle responsivo ao tráfego (TR), que consiste em projetar múltiplos planos de tempo e estratégias para comutar os planos em resposta às condições do tráfego prevalecentes; e (ii) o controle adaptativo ao tráfego (TA), que continuamente examina o tráfego corrente, prevê as suas condições, e ajusta os sinais de controle baseado nas condições detectadas de forma a otimizar o seu desempenho. Estudos mostram que o controle TA é altamente complexo, incorre custos altos em relação aos orçamentos das cidades e depende do conhecimento e intervenção freqüente de profissionais especializados.

Recentemente, sistemas multiagentes e tecnologia de aprendizagem de máquina têm sido empregados para contornar a complexidade inerente aos sistemas e torná-los mais autônomos. Thorpe e Anderson (Thorpe e Anderson, 1996) aplicaram o algoritmo SARSA para controlar os semáforos de uma rede pequena, mostrando que o tempo de espera nos semáforos poderia ser reduzido na proporção de 87% se comparados a uma estratégia de controle padrão. Mais recentemente, Wiering (Wiering, 2000) desenvolveu algoritmos baseados em aprendizagem por reforço para controlar os semáforos—i.e., um modelo foi usado para calcular a dinâmica do tráfego e então um método semelhante a programação dinâmica foi aplicado para computar funções valor-estado, as quais quantificam os ganhos acumulados por um agente a longo prazo a partir dos diversos estados do sistema. A pesquisa a ser relatada daqui por diante pode ser vista como uma extensão a estas investigações pioneiras.

Este artigo estrutura o problema de operação de uma rede de tráfego veicular como um jogo estocástico entre agentes de controle distribuídos, provendo desse modo uma nova perspectiva do problema e um meio de se entender a interação entre os agentes. Ele enfoca o desenvolvimento de estratégias de controle TR e propõe um algoritmo de aprendizagem livre de modelo (*model-free*), denominado *Q-learning*, para aprender estratégias de controle otimizadas. Mais adiante, se obtêm evidências experimentais do mérito do algoritmo proposto a partir da simulação de uma rede de tráfego veicular representativa.

2. CONTROLE E GERENCIAMENTO DE REDES DE TRÁFEGO

Para se apreciar o aumento na qualidade da operação de redes de tráfego veicular impulsionado pelo emprego de tecnologias de controle e otimização modernas, torna-se necessário o entendimento das diferenças entre os tipos de controle de semáforos, os quais podem ser divididos em três categorias principais (Garbacz, 2002):

Controle pela hora do dia (TOD): Refere-se ao controle semafórico coordenado com base em planos de tempo fixo, os quais são projetados *off-line* por engenheiros de tráfego ou definidos por software de propósito específico e posteriormente implantados no sistema de controle de tráfego. Um software extensamente usado para computar planos de tempo fixo é o Transyt (Crabtree, Vincent e Harrison, 1996), o qual tem servido de padrão para comparação com políticas de controle sintetizadas a partir de planos de tempo fixo. A eficácia do controle TOD se apóia no padrão cíclico do tráfego e varia com base na hora do dia e dia da semana. A partir das estimativas horárias de chegada de veículos, o comprimento do ciclo e a fração de tempo de verde alocada a cada fila são computados para cada semáforo. A eficácia do controle TOD depende da suposição que os padrões de tráfego evoluem gradualmente com o passar dos anos. Todavia, os padrões de tráfego podem mudar drasticamente em resposta a condições inesperadas. Devido às fortes suposições sob as quais o controle TOD está fundamentado, não se espera um aumento

substancial no desempenho das redes de tráfego por meio do simples aprimoramento do projeto de planos de tempo fixo.

Controle responsivo ao tráfego (TR): Esta categoria compreende as estratégias de controle que implementam planos de tempo fixo com coordenação pré-programada, mas ao invés de seguir uma programação fixa, ela seleciona planos conforme as condições de tráfego prevalecentes. Para sua implementação, sensores veiculares localizados próximo aos cruzamentos fornecem um fluxo contínuo de dados sobre o tráfego em tempo real para o controlador, que comuta para outro plano de tempo fixo quando o nível de tráfego alcança limites pré-definidos. A pré-computação de planos de tempo fixo, a combinação com condições de tráfego estimadas a partir de dados sensoriais e a seleção do melhor plano para as condições vigentes do tráfego estão longe de serem tarefas fáceis. Em uma escala menor, o controle TR sofre das mesmas deficiências do controle TOD, no sentido de que ambos confiam no planejamento *off-line* que pode se tornar ineficiente à medida que os padrões de tráfego evoluem.

Controle adaptativo ao tráfego (TA): O controle adaptativo ao tráfego ajusta os parâmetros de tempo dos semáforos continuamente a fim de obter uma melhora nas métricas de desempenho em resposta às condições de tráfego passadas, atuais e antecipadas. Assim, o controle TA oferece os benefícios de lidar com padrões de tráfego incomuns ou não planejados, reduzindo paradas e atrasos sob condições normais, ao mesmo tempo que fornece um meio de se adaptar a padrões de tráfego variados. Mas, contrário ao que muitos acreditam, o controle de tráfego adaptativo não requer menos conhecimento especializado e manutenção que o controle de tempo fixo convencional. Uma vez que o controle TA é inerentemente complexo, sua eficácia de implementação é mais uma arte que uma ciência, dependendo intrinsecamente do conhecimento de engenheiros de tráfego experientes. Três sistemas de controle TA representativos são: Sistema de Tráfego Adaptável Coordenado de Sydney (SCATS) (Abdel-Rahim, Taylor e Bangia, 1998), a Técnica de Otimização Compensatória duração/comprimento de Ciclo Dividido (SCOOT) (Robertson e Bretherton, 1991) e a estratégia de controle de Políticas Otimizadas para Controle Adaptativo (OPAC) (Gartner, Pooran e Andrews, 2001).

A partir da avaliação acima, podemos deduzir que a tecnologia de controle de tráfego é aplicada em graus variados, com abordagens em malha-aberta e malha-fechada, cada uma tendo suas vantagens e deficiências. Em uma extremidade, o controle TOD é sem dúvida o menos oneroso e tem potencial para induzir uma operação próxima do ideal em certas situações, tais como ao longo de rodovias arteriais cujos padrões de tráfego são previsíveis. No meio da escala, o controle TR pode produzir desempenho superior ao controle TOD, em particular quando os padrões de tráfego variam com o tempo, mas este necessita de revisão periódica, da intervenção de peritos e de uma rede de sensores na vizinhança dos cruzamentos. Na outra extremidade, o controle TA é o mais promissor quanto ao aumento do desempenho, mas a sua implantação e manutenção têm custos elevados e são de natureza complexa, além de exigir conhecimento especializado.

3. AGENTES DISTRIBUÍDOS DE APRENDIZAGEM POR REFORÇO

Como salientou Garbacz (Garbacz, 2002), os custos de manutenção do controle adaptativo ao tráfego são substancialmente mais vultosos que os custos de manutenção de estratégias

convencionais de controle, o que pode ser ainda mais desfavorável se considerarmos a dependência de especialistas em controle de tráfego. Isto significa que os benefícios promovidos pelo controle TA podem não justificar os custos de sua implantação. A complexidade pode ser ainda mais exacerbada se os processos de tomada de decisão são distribuídos sobre uma área geográfica (Gartner, Pooran e Andrews, 2001), mesmo quando métodos de otimização sofisticados forem empregados, tais como a programação dinâmica e a abordagem de horizonte deslizante. Devido a esses obstáculos, parece que ferramentas semiautomáticas deveriam ser desenvolvidas para tornar os sistemas de controle existentes mais autônomos. Para este fim, técnicas de aprendizagem por reforço (Kaelbling, Littman e Moore, 1996) ou programação dinâmica baseada em simulação (Bertsekas, 1995), como é conhecida dentro da teoria de controle ótimo, detêm o potencial de colocar à disposição o grau requerido de autonomia. Em primeiro lugar, a fundamentação matemática destes métodos está centrada nos processos de decisão Markovianos e na programação dinâmica, que fornece condições teóricas de desempenho ótimo e que serve de padrão contra o qual outros métodos, talvez mais práticos, possam ser comparados. Métodos alternativos, tal como a programação neuro-dinâmica, tendem a aproximar a função custo remanescente da programação dinâmica (i.e., as funções valor-estado e valor-ação). Outro aspecto relevante é o sucesso obtido por técnicas de aprendizagem por reforço em problemas desafiadores, em particular na manipulação robótica e em problemas de navegação, que indica a possibilidade de se estender tais técnicas para o domínio do controle de tráfego veicular.

O problema padrão de aprendizagem por reforço (i.e., o problema de aprender uma política de controle que maximize a recompensa amortizada de um agente que atravessa um processo de decisão Markoviano) difere do problema de controle de redes de tráfego veicular em pelo menos dois aspectos: (i) o controle de tráfego é o esforço coletivo de centenas de dispositivos de controle distribuídos cujas políticas de controle ótimas dependem das políticas implementadas pelos demais agentes; e (ii) o estado da rede é observado parcialmente pelos dispositivos de controle. Portanto, o problema de controle da rede de tráfego pode ser modelado como um jogo estocástico distribuído (Camponogara e Kraus Jr, 2003) dado por:

- N: número de agentes de controle ou dispositivos distribuídos;
- S: conjunto discreto dos possíveis estados da rede de tráfego, consistindo no número de veículos nas vias;
- $\theta = \{\theta_n\}$: conjunto de funções que modela a visão parcial dos agentes da rede, onde $\theta_n(s)$ é a fração do estado $s \in S$ percebida pelo agente n; no caso de semáforos, θ_n proveria informações relativas apenas à vizinhança do semáforo sob controle do agente n;
- $A = A_1 \times \ldots \times A_N$: conjunto de ações, ou seja, sinais de controle de tráfego, onde A_n é o subconjunto de ações delegadas ao agente n;
- $T : S \times A \times H \times S \to [0,1]$: função de transição de estados que simula a dinâmica do fluxo de tráfego, onde $T(s,a,h,s')$ é a probabilidade de se alcançar o estado s' a partir do estado s se a ação de controle a é tomada no instante h, o que implica $\sum_{s' \in S} T(s,a,s',h) = 1$, $\forall s \in S$, $\forall a \in A$ e $\forall h \in H$; freqüentemente, a função de transição de estados é aproximada por meio de um simulador; e
- $R = \{R_n\}$: conjunto de sinais de reforço, onde $R_n : S \times A \times S \to \mathbb{R}$ fornece o sinal de recompensa para o agente n correspondente à transição do estado s_t para s_{t+1} se a ação

comum a for executada; R_n define claramente o comportamento desejado para um agente n, que pode ser ajustado para se reduzir atrasos nas interseções ou aumentar o fluxo de tráfego.

De forma mais compacta, um jogo estocástico distribuído pode ser representado por uma tupla $\Gamma = (N, S, \theta, A, H, T, R)$ que agrega os elementos listados acima. Pode-se pensar sobre jogos estocásticos como generalizações de processos de decisão Markoviano (no sentido que os múltiplos jogadores tomam decisões independentes e recebem recompensas que dependem das decisões conjuntas e das transições de estado) e modelos da teoria de jogos (no sentido que os jogos estocásticos podem ser vistos como uma série de jogos que evoluem através do espaço de estados). Deste modo, a extensão do problema de aprendizagem por reforço de um único agente para jogos estocásticos tem como obstáculo a dificuldade em se encontrar um conjunto de políticas $\pi = (\pi_1, \ldots, \pi_N)$ tal que cada política π_n seja ótima do ponto de vista do agente n. Uma política π_n é uma função $\pi_n : S_n \times A_n \rightarrow [0,1]$ tal que $\pi_n(s_n, a_n)$ corresponde à probabilidade do agente n tomar a ação a_n quando este se encontrar no estado s_n.

No presente estudo, entretanto, a noção de otimalidade tem que ser melhor elaborada. Porque cada agente faz o melhor para si próprio, a política do agente só pode ser ótima se o conjunto de todas as políticas induz um equilíbrio, isto é, um ponto de equilíbrio Nash (Basar e Olsder, 1999). O conjunto de políticas π produz um equilíbrio Nash, no sentido estocástico, se cada agente n não tiver qualquer incentivo para divergir de seu política de decisão enquanto os outros agentes se mantiverem estáveis quanto às suas políticas atuais. Esta questão tem sido objeto de estudos teóricos e práticos, tais como a existência de pontos de equilíbrio Nash, a convergência das políticas dos agentes para atratores, a velocidade de convergência produzida por algoritmos de busca de políticas e por último, mas não menos relevante, a qualidade da operação induzida pelos agentes de tomada de decisão. Em princípio, se recursos computacionais ilimitados estivessem disponíveis, uma operação ótima (i.e., um conjunto de decisões Pareto ótimas) poderia ser obtida com um agente centralizado. Ao contrário de processos de decisão Markoviano, nem a existência e nem a convergência para pontos de equilíbrio Nash podem ser garantidos em jogos estocásticos. Conseqüentemente, na prática, estas questões são abordadas caso a caso por meio de análise computacional e simulações.

Jogos estocásticos se tornaram o formalismo preferido para modelar sistemas multiagentes (Bowling e Veloso, 2002). A distinção principal entre jogos estocásticos distribuídos e sua forma padrão encontra-se na visão limitada do estado por parte dos agentes distribuídos, isto é, cada agente n detecta o valor de apenas uma fração das variáveis, a saber $\theta_n(s)$.

Pelas razões explicitadas acima, propomos a aplicação de modificações de algoritmos para aprendizagem por reforço de um único agente ao problema de encontrar um conjunto de políticas de controle ótimas π. Com relação às três classes principais de algoritmos de aprendizagem por reforço, os métodos de diferença temporal destacam-se como os mais promissores. As evidências experimentais e científicas obtidas até o momento demonstram que os métodos de diferença temporal podem ser eficazes: os algoritmos de programação dinâmica exigem informação precisa da função de probabilidade de transição de estado, T; e a taxa de convergência dos métodos tipo Monte Carlo são excessivamente lentas; por outro lado, a predição e aprendizagem por diferença temporal tem sido muito próspera em uma variedade de aplicações, incluindo problemas de otimização e controle. Para os desenvolvimentos daqui em diante, $\lambda_n = \pi - \{\pi_n\}$ denotará o conjunto das políticas dos agentes a exceção do agente n. Assumindo um conjunto de políticas estacionárias λ_n, a função ação-valor induzida pela

política π_n do agente n, para o estado $s_n \in S_n$ onde $S_n = \{s_n : s_n = \theta_n(s)$ para algum $s \in S\}$, e a ação $a_n \in A_n$, é definida como:

$$Q_n^{\pi_n,\lambda_n}(s_n,a_n) \;=\; E_{\pi_n,\lambda_n}[\sum_{k=0}^{L}\gamma^k r_{n,t+k+1} : s_{n,t}=s_n, a_{n,t}=a_n] \qquad (1)$$

onde L é a duração do episódio, $0 \le \gamma \le 1$ é o fator de amortização e $r_{n,t+1}$ é o ganho recebido pelo agente n ao executar a ação a_n no instante t e causar a transição do estado $s_{n,t} = s_n$ para $s_{n,t+1}$. Em outras palavras, $Q_n(s_n,a_n)$ é o ganho amortizado esperado que o agente n recebe, se este começa no estado s_n, toma a ação a_n e depois segue a política π_n, enquanto os agentes restantes se comportam de acordo com λ_n. O episódio pode ser qualquer ciclo periódico tais como dias e semanas. A fim de acumular seu ganho amortizado máximo, o agente n busca aprender uma função valor-ação ótima dada por:

$$Q_n^{\pi_n^*,\lambda_n}(s_n,a_n) \;=\; \begin{array}{ll}\text{Maximize} & Q_n^{\pi_n,\lambda_n}(s_n,a_n) \\ \pi_n & \text{enquanto } \lambda_n \text{ é invariante}\end{array} \qquad (2)$$

para cada $s_n \in S_n$ e $a_n \in A_n$. Do ponto de vista do agente n, a política ótima é uma função das políticas dos outros agentes: $\pi_n^* = \Pi_n^*(\lambda_n)$. Isto dá origem à questão da existência de um conjunto de políticas π^* que sejam simultaneamente ótimas para cada agente n, isto é, $\pi_n^* = \Pi_n^*(\lambda_n^*)$ para cada n onde $\lambda_n^* = \pi^* - \{\pi_n^*\}$. Tal conjunto π^* é conhecido como conjunto de políticas de equilíbrio Nash (Basar e Olsder, 1999): nenhum agente racional n divergirá de π_n^* porque caso contrário este incorreria perdas a si próprio. As políticas de equilíbrio Nash induzem funções valor-ação ótimas a todos os agentes, as quais podem ser expressas como:

$$\left\{ \begin{array}{lll} Q_n^{\pi_n^*,\lambda_n^*}(s_n,a_n) \;=\; \begin{array}{ll}\text{Maximize} & Q_n^{\pi_n,\lambda_n^*}(s_n,a_n) \\ \pi_n & \text{enquanto } \lambda_n^* \text{ é invariante}\end{array} & & \text{para } n = 1,\ldots,N \end{array} \right. \qquad (3)$$

A existência de políticas de equilíbrio Nash e a convergência para tais políticas são temas recorrentes no campo dos jogos estocásticos, para os quais não existem respostas definitivas. Freqüentemente, estas questões são analisadas caso a caso por meio de experimentação numérica. Para aproximar políticas Nash, propomos a busca iterativa por um conjunto ótimo de funções valor-ação, como descrito em (3), pela qual a operação da rede sob a política dos agentes é iterativamente simulada e melhorada ao fim de cada episódio. Esta estratégia de busca consiste na aplicação de um algoritmo *Q-learning* modificado para cada agente, daqui por diante denominado algoritmo *Q-learning* distribuído, cujo pseudo-código é dado abaixo.

Algoritmo *Q-Learning* Distribuído

```
Cada agente n inicializa Q_n(s_n, a_n) arbitrariamente
Repita (para cada episódio)
   Inicializa s
   Repita para cada passo do episódio
      Para cada agente n faça
        Escolha uma ação a_n com base em Q_n(s_n), s_n = θ_n(s),
           usando uma política derivada de Q_n tal como ε-gulosa
        Implemente a ação a_n
      Fim-para
```

```
      Aguarde a rede reagir às ações e evoluir do estado s para s'

      Para cada agente n faça
        Observe r_n e s'_n = θ_n(s'), onde s' é o próximo estado
        Q_n(s_n,a_n) ← Q_n(s_n,a_n) + α_n[r_n + γ_n max_{a'_n} Q_n(s'_n,a'_n) − Q_n(s_n,a_n)],
            onde α_n ∈ [0,1] é a taxa de aprendizagem e
            γ_n ∈ [0,1] é a taxa de amortização do agente n
      Fim-para
   Fim-repita
Fim-repita
```

4. ANÁLISE COMPUTACIONAL

Esta seção avalia o potencial do algoritmo *Q-learning* para a síntese de políticas de controle de tráfego responsivas em uma seção pequena, mas representativa da rede de tráfego da cidade de Florianópolis. A Figura 1 descreve a sub-rede que se estende ao longo da Avenida Beira-Mar. A sub-rede, representada por Ω, foi modelada como um jogo estocástico distribuído Γ cujos elementos são:

- $N = 13$ é o número de agentes controladores, indicados na Figura da rede de interesse por I_n.
- S é o conjunto de estados obtido ao agregar-se o número de veículos parados em cada uma das vias que conduzem a uma das interseções I_n, mas este número é aproximado com valores $0, 4, 8, \ldots, 64$ para manter o uso da memória sob limites;
- θ_n mapeia $s \in S$ para o sub-conjunto de variáveis que correspondem ao número e a posição dos veículos nas vias adjacentes à interseção I_n, ou seja, θ_n retorna ao agente n somente o estado das vias que conduzem à interseção I_n;
- A_n modela o conjunto de ações de controle nos semáforos, indicando os movimentos de tráfego factíveis que podem ser executados simultaneamente numa interseção I_n;
- T é a função de transição de estados, a qual segue os padrões de tráfego delineados acima, tendo sido obtida a partir de estatísticas geradas com base no fluxo de tráfego veicular nas interseções, estatísticas essas fornecidas pelo Instituto de Planejamento Urbano de Florianópolis (IPUF); e
- R é um conjunto de funções sinal-reforço, onde R_n retorna a cada passo de simulação o valor negativo do número de veículos aguardando passagem pela interseção I_n e que podem ser observados pelo agente n.

A evolução do estado da rede de tráfego Ω com o passar do tempo, em resposta à dinâmica do fluxo de tráfego e aos sinais de controle, foi aproximada com um simulador desenvolvido previamente (Camponogara e Kraus Jr, 2003). Este simulador foi feito sob medida para simular jogos estocásticos que modelam redes de tráfego veicular e que venham a servir de base para investigação computacional, foi implementado em linguagem C/C++ ANSI para incentivar seu

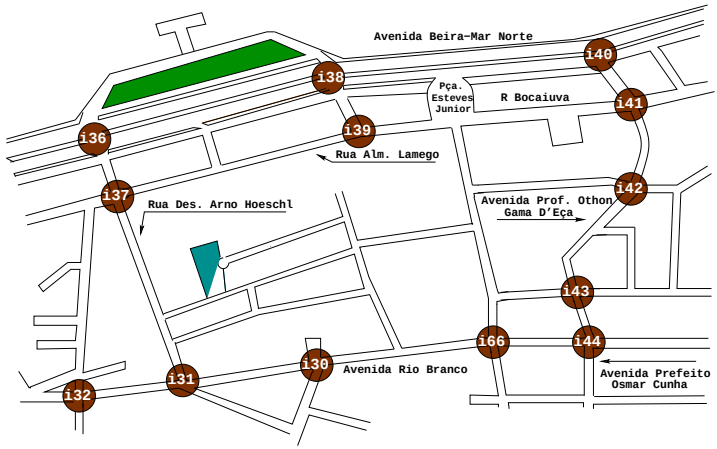

Figura 1: A sub-rede de interesse da cidade de Florianópolis

uso em computadores baseados em Unix e, igualmente importante, para facilitar a integração com algoritmos de controle semafórico. Pode-se com facilidade modelar Ω como um grafo dirigido $G = (V,E)$, cujos nós representam as interseções e cujos arcos representam as vias. O grafo G e todos os outros elementos de Γ são definidos em arquivos tipo texto que são posteriormente carregados pelo simulador. Com a finalidade de comparação, serão adotadas três estratégias de controle para decidir os sinais semafóricos:

Política Aleatória Uniforme: atribui uma probabilidade igual a cada conjunto de movimentos simultaneamente realizáveis, isto é, $\pi_n(s_n, a_n) = \frac{1}{|A_n|}$ para cada agente $n \in \{1, \ldots, N\}$, estado $s_n \in S_n$ e ação de controle $a_n \in A_n$. Esta estratégia, talvez a mais simples dentre todas as estratégias, é insensível às condições de tráfego.

Política de Controle de Melhor-Esforço: atribui a cada conjunto de movimentos que podem ser realizados simultaneamente, isto é, a cada elemento de A_n, uma probabilidade proporcional ao número de veículos parados que podem progredir se a_n for executado. Contrária à estratégia anterior, esta é sensível às condições de tráfego vigentes já que tende a favorecer o fluxo das filas mais longas, ou seja, ela libera o fluxo de tráfego para a via que apresenta um maior número de veículos em espera.

Política *Q-Learning* Distribuída: os agentes de controle buscam um conjunto de políticas de controle π interagindo com o modelo da rede, procurando iterativamente por funções valor-ação ótimas. Desse modo o conjunto de políticas de controle π sintetizada pelos agentes tende a se aproximar da política centralizada ótima π^*, a qual maximiza a soma dos ganhos de todos os agentes a longo prazo, o que se entende como objetivo principal da operação da rede de tráfego veicular. Nos experimentos computacionais variou-se o número de agentes que empregam a política *Q-learning*, enquanto que a política aleatória uniforme foi implementada nas interseções restantes. A taxa de amortização foi fixada em $\gamma_n = 0.9$ e a taxa de aprendizado foi $\alpha_n = 0.1$ para cada agente n.

Os resultados dos experimentos numéricos como delineados acima são relatados na Tabela 1. Onde a coluna mais à esquerda se refere à densidade de tráfego (σ), parâmetro este que descreve

Tabela 1: Número médio de veículos em espera para densidades de tráfego e políticas de controle diversas

σ	Número médio de veículos em espera								Porcentagem de Ganho	
	Aleatória Uniforme	Melhor Esforço	Número de agentes *Q*-learning							
			2	4	6	8	10	13	Aleatória	Melhor
0.05	59.84	12.83	46.43	32.05	20.86	13.84	8.72	4.59	92.32	64.22
0.10	123.88	26.51	92.93	67.75	43.36	29.72	18.99	10.87	91.22	58.99
0.15	210.09	57.51	151.80	117.79	88.61	71.01	60.27	38.83	81.51	32.48
0.20	378.43	229.35	242.48	206.53	171.77	150.21	132.16	125.89	66.73	45.11
0.25	613.28	395.00	445.92	376.20	327.47	316.01	289.11	249.15	59.37	36.92
0.30	849.66	620.04	627.08	531.64	452.58	445.80	414.79	337.52	60.27	45.56
0.35	981.18	858.88	755.79	646.24	555.35	477.37	521.27	452.29	53.90	47.33
0.40	1201.18	1065.05	976.56	794.15	708.15	714.07	654.81	612.25	49.02	42.51
0.45	1269.70	1162.32	1058.22	898.89	787.98	755.92	789.29	668.12	47.37	42.51
0.50	1510.06	1299.54	1249.17	1042.54	935.48	873.46	862.03	841.22	44.29	35.26
0.55	1651.67	1540.60	1432.13	1249.38	1072.99	1048.24	1023.17	992.83	39.88	35.55
0.60	2018.43	1739.25	1776.56	1509.69	1421.95	1242.47	1339.25	1245.49	38.29	28.38
0.65	2332.67	2011.49	2160.62	1855.77	1688.94	1747.05	1821.55	1455.51	37.60	27.64
0.70	2667.03	2213.37	2405.14	2218.41	2245.10	2128.22	2122.92	1900.03	28.75	14.15
0.75	2975.95	2515.14	2800.97	2531.15	2509.55	2229.26	2289.89	2161.41	27.37	14.06
Média	1256.20	1049.79	1081.45	938.54	868.68	816.18	823.21	739.73	41.11	38.04
Ciclo Total	1181.76	995.80	1066.33	895.84	773.71	740.73	725.61	738.45	30.75	25.85

o nível de ocupação de Ω, que pode ser variado de forma a modelar desde uma rede vazia ($\sigma = 0$) a uma rede com indíce de ocupação máximo ($\sigma = 1.0$). O parâmetro de densidade de tráfego pode ser usado para simular condições de tráfego diversas. As outras colunas dão a média do número total de veículos em espera nas interseções induzidas pela política aleatória uniforme, pela política de controle de melhor-esforço, e pela política *Q-learning* distribuída para um número variado de agentes baseados em aprendizagem por reforço. As médias foram tomadas durante 10 episódios, cada um consistindo de 2000 passos de iteração, significando que o horizonte de tempo é $L = 2000$. Para o caso do *Q-learning* distribuído, os agentes foram sujeitados a um período de 600 episódios de aprendizagem antes de avaliar o desempenho da rede. As duas últimas colunas mostram o ganho em desempenho produzido por 13 agentes *Q-learning* com respeito ao desempenho produzido pelas políticas de controle aleatória uniforme e de melhor-esforço.

Objetivando aferir o desempenho relativo das políticas de controle sob condições de tráfego variáveis no tempo (e.g., σ variável), conduzimos uma série de experiências computacionais nas quais a densidade de tráfego variou durante cada episódio: σ assumiu valores do conjunto $\{0.25, 0.40, 0.55, 0.70\}$ e os episódios duraram 5000 passos de tempo. Os resultados destas experiências são descritos na linha inferior da Tabela 1. A Tabela relata o número médio dos veículos em espera induzido pelas políticas de controle aleatória uniforme, de controle de melhor-esforço e de controle distribuído para números variados de agentes. Onde, na Figura 2 são ilustrados os desempenhos dos métodos da política aleatória uniforme e do *Q-learning* adotado por 13 agentes para a rede com densidade variada, segundo a trajetória dos veículos em espera. A evidência numérica coletada a partir destas experiências computacionais confirma a hipótese de que as técnicas de aprendizagem de máquina podem produzir ganhos significativos na operação de redes de tráfego veicular.

A sub-rede descrita na Figura 1, Ω, foi modelada aproximadamente no simulador *Green Light District* (GLD) (Wiering, 2000) e, subseqüentemente, experiências foram conduzidas

com agentes de aprendizagem por reforço. O propósito desta análise foi medir o ganho em desempenho, obtido pelo uso de técnicas de aprendizagem de máquina no controle de semáforos, em condições onde os veículos são modelados como agentes individuais. Nossa análise consistiu em usar a plataforma de simulação de agentes baseados em aprendizagem por reforço desenvolvida por Wiering (Wiering, 2000) para operar a sub-rede da Figura 1. Uma breve descrição dos resultados destas experiências é dada na Tabela 2. Onde a coluna mais à esquerda descreve a densidade de fluxo de tráfego na sub-rede. As três colunas seguintes mostram a média do comprimento das filas de veículos em espera nas interseções durante dez simulações, enquanto as duas últimas colunas mostram o ganho em desempenho obtido pelos agentes de controle baseados em aprendizagem por reforço em comparação aos desempenhos obtidos pelas políticas aleatória uniforme e de fila mais longa.

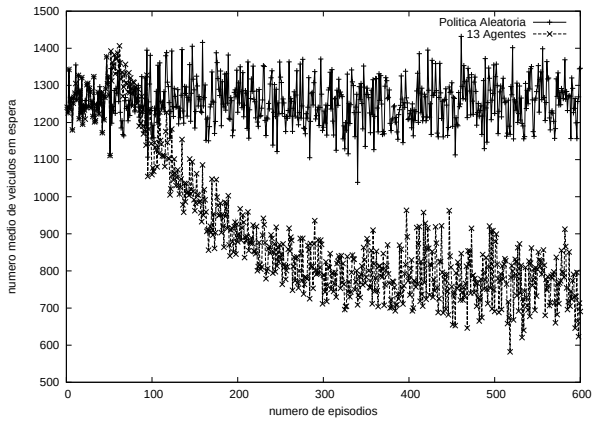

Figura 2: Desempenho dos métodos *Q-learning* para 13 agentes e da política aleatória uniforme sob densidade de tráfego variada

5. CONSIDERAÇÕES FINAIS

Para atender às demandas dos ecologistas, motoristas e usuários, avanços em infra-estrutura e tecnologia de controle deverão ser implementados nos sistemas de tráfego, mas individualmente estes avanços não serão eficientes e eficazes em particular no que se refere às questões econômicas. As duas alternativas ao controle de planos fixos (TOD), o controle responsivo (TR) e o controle adaptativo (TA), possuem altos custos, necessitam de dispositivos sensoriais, são de complexidade considerável e dependem da intervenção periódica de engenheiros especialistas. Para este fim, este artigo revisou brevemente as tecnologias de controle de tráfego existentes, sugerindo uma nova perspectiva para o problema de controle otimizado de redes de tráfego por meio da formalização do problema como um jogo estocástico entre agentes de controle distribuídos. Foram apresentadas noções de equilíbrio, discutidas as condições de otimalidade das políticas de controle distribuídas e um algoritmo de aprendizagem distribuído foi sugerido para aproximar políticas ótimas. Os resultados computacionais, obtidos a partir de simulações de uma sub-rede representativa, indicam que a tecnologia de aprendizagem de máquina pode ser uma alternativa eficaz para contornar a complexidade da operação de redes de tráfego enquanto, ao mesmo tempo, proporciona ganhos em desempenho. É nossa intenção avançar esta tecnologia e incorporá-la a sistemas de controle de tráfego protótipo e até mesmo comerciais.

Tabela 2: Comprimento médio da fila em espera

	Comprimento médio da fila			Porcentagem Ganho	
σ	Aleatória uniforme	Fila mais longa	Agentes baseados em AR	Aleatória	Fila maior
0.05	774	1291	374	51.67	71.03
0.25	1971	2244	578	70.67	74.24
0.50	2241	2244	2217	1.07	1.20
0.75	2239	2244	2196	1.92	2.13
Média	1806	2005	1341	31.33	37.15

Agradecimentos

Os autores agradecem ao CNPq/SEPIN/FINEP pelo suporte à pesquisa recebido através do financiamento ao projeto SincMobil (processo 552248/02-9) e à CAPES pela concessão de uma bolsa de mestrado ao segundo autor.

REFERÊNCIAS BIBLIOGRÁFICAS

Abdel-Rahim, A.; W. Taylor; A. Bangia (1998) The impact of SCATS on travel time and delay. In: *Proceedings of the 8th Annual Meeting of the Intelligent Transportation Society of America*. Detroit, MI, U.S.A.

Basar, T.; G. Olsder (1999) *Dynamic Noncooperative Game Theory*. 2nd. ed. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Bertsekas, D. (1995) *Dynamic Programming and Optimal Control*. 2nd. ed. Belmont, Massachusetts: Athena Scientific.

Bowling, M.; M. Veloso (2002) Scalable learning in stochastic games. In: *Proceedings of the AAAI Workshop on Game Theoretic and Decision Theoretic Agents*. Edmonton, Alberta, Canada: The AAAI Press.

Camponogara, E.; W. Kraus Jr (2003) Distributed learning agents in urban traffic control. *Proceedings of the 11th Portuguese Conference on Artificial Intelligence*. Beja, Portugal: Springer-Verlag. (Lecture Notes in Artificial Intelligence, v. 2902), p. 324–335.

Crabtree, M.; R. Vincent; S. Harrison (1996) *TRANSYT: 10 user's guide*. Crawthorne, UK.

Garbacz, R. (2002) Adaptive signal control: what to expect. In: *Proceedings of the 12th Annual Meeting of the Intelligent Transportation Society of America*. Washington, DC, U.S.A.

Gartner, N.; F. Pooran; C. Andrews (2001) Implementation of the OPAC adaptive control strategy in a traffic signal network. In: *Proceedings of the IEEE 4th International Conference on Intelligent Transportation Systems*. Oakland, U.S.A.: IEEE Press. p. 195–200.

Hunt, P. et al. (1981) *SCOOT: a traffic responsive method of coordinating signals*. Crowthorne, England.

Kaelbling, L.; M. Littman; A. Moore (1996) Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, v. 4, p. 237–285, February.

Robertson, D.; R. Bretherton (1991) Optimizing networks of traffic signals in real time: the SCOOT method. *IEEE Transactions on Vehicular Technology*, v. 40, n. 1, p. 11–15, February.

Thorpe, T.; C. Anderson (1996) *Traffic light control using SARSA with three state representations*. Boulder, CO, U.S.A.

Wiering, M. (2000) Multi-agent reinforcement learning for traffic light control. In: *Proceedings of the 17th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann. p. 1151–1158.