MODELO HÍBRIDO DE APRENDIZADO POR REFORÇO E LÓGICA FUZZY APLICADO AO PROBLEMA DO CAIXEIRO VIAJANTE COM REABASTECIMENTO

André Luiz Carvalho Ottoni*, Erivelton Geraldo Nepomuceno*, Marcos Santos de Oliveira*, Leonardo Bonato Felix*

* *Programa de Pós-Graduação em Engenharia Elétrica (UFSJ/CEFET-MG)*
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil

Emails: andreottoni@ymail.com, nepomuceno@ufsj.edu.br, mso@ufsj.edu.br, leobonato@ufv.br

Abstract— This work has as main objective to develop a hybrid model of Reinforcement Learning and Fuzzy Logic for application in the Traveling Salesman Problem with Refueling. For this, were implemented four RL methods, with two of those with a fuzzy reinforcement function. The methods with fuzzy reward, Q-Fuzzy and S-Fuzzy, showed the best results.

Keywords— Reinforcement Learning, Fuzzy Logic, Traveling Salesman Problem.

Resumo— Este trabalho tem como objetivo principal desenvolver um modelo híbrido de Aprendizado por Reforço e Lógica Fuzzy para aplicação no Problema do Caixeiro Viajante com Reabastecimento. Dessa forma, foram implementados quatro métodos de AR, sendo que dois desses com uma função de recompensa fuzzy. Os métodos com recompensa fuzzy, Q-Fuzzy e S-Fuzzy, apresentaram os melhores resultados.

Palavras-chave— Aprendizado por Reforço, Lógica Fuzzy, Problema do Caixeiro Viajante.

1 Introdução

A administração nos custos no transporte é fundamental nos dias atuais. No Brasil, o custo com combustível é alto, e representa uma parcela significativa nos gastos de uma transportadora (Rodrigues Junior, 2011). Nesse aspecto, vale ressaltar que existe uma variação importante entre os preços dos combustíveis em uma mesma região rodoviária (Rodrigues Junior, 2011). Esse fato pode ser verificado com pesquisas de preços de combustíveis no site da Agência Nacional de Petróleo (ANP)[1].

Os modelos de reabastecimento buscam otimizar o gasto com combustíveis. Segundo Khuller et al. (2007), existem quatro importantes tipos de problemas de reabastecimento: 1) Reabastecimento com rota fixa, 2) Reabastecimento com rota variável, 3) Problema do Caixeiro Viajante (PCV) considerando custo uniforme em cada ponto, 4) PCV com o custo do combustível variando em cada ponto.

Na última década, alguns trabalhos abordaram e desenvolveram métodos para problema de reabastecimento de veículos. O artigo de Lin et al. (2007) desenvolveu um algoritmo de tempo de execução linear para encontrar a política ótima de reabastecimento para um problema de rota fixa. Já o trabalho de Suzuki (2008), apresenta um modelo que leva em consideração os custos operacionais do veículo ao longo da trajetória. Em Suzuki (2012), por sua vez, aborda o PCV com janela de tempo e reabastecimento. Já a pesquisa de Rodrigues Junior e Cruz (2013), utiliza uma modelagem com rotas fixas para otimizar um estudo de caso real de uma transportadora. Os autores de Silva e Cruz (2014) apresentam um estudo referencial teórico sobre alguns modelos de otimização para o problema de reabastecimento de veículos.

Um método abordado para a resolução de problemas de roteamento, como o Problema do Caixeiro Viajante, é o Aprendizado por Reforço (AR) (Lima Júnior et al., 2010; Ottoni et al., 2015). O Aprendizado por Reforço é uma técnica de aprendizado de máquina modelada por Processos de Decisão de Markov (Watkins e Dayan, 1992; Sutton e Barto, 1998). No AR, o aprendizado se baseia nos resultados de sucesso e fracasso de um agente interagindo em um ambiente. O agente aprende a tomar a decisão nos estados do ambiente, recebendo recompensas por ações corretas e penalidades por ações erradas.

Uma das vertentes dos estudos de aprendizado por reforço são os modelos híbridos. Nesses modelos, outras técnicas inteligentes são aplicadas em conjunto com AR na tentativa de melhorar o desempenho do sistema (Bianchi, 2004). Nessa perspectiva algumas pesquisas adotam algoritmos de aprendizado por reforço em conjunto com lógica fuzzy. O trabalho Zhou e Meng (2003) apresenta um método *fuzzy reinforcement learning* (FRL) para o controle de um robô bípede. Em Faria e Romero (2002) por sua vez, propõe o método R'-learning, que incorpora lógica fuzzy ao algoritmo R-learning. Já os trabalhos de Bonarini et al. (2009) e Derhami et al. (2010), levantam aspectos

[1] http://anp.gov.br/preco/

da aplicação fuzzy em conjunto com os algoritmos *Q-learning* e SARSA, respectivamente.

Baseando-se nisso, este trabalho tem como objetivo principal desenvolver um modelo híbrido de Aprendizado por Reforço e Lógica Fuzzy para aplicação no Problema do Caixeiro Viajante com Reabastecimento. Dessa forma, foram implementados quatro métodos de AR, sendo que dois desses com uma função de recompensa fuzzy.

Este artigo está organizado em seções. Na seção 2, são definidos conceitos teóricos do aprendizado por reforço. Na seção 3, o problema do caixeiro viajante com reabastecimento é descrito. Já na seção 4, é apresentada a modelagem do sistema híbrido. Uma descrição dos experimentos realizados é apresentada na seção 5. A análise dos resultados obtidos é apresentada na seção 6. Finalmente, na seção 7 são apresentadas as conclusões.

2 Aprendizado por Reforço

Alguns algoritmos de AR são frequentemente adotados na literatura, como o *Q-learning* (Watkins e Dayan, 1992) e o SARSA (Sutton e Barto, 1998), apresentados em seguida.

O *Q-learning* (Watkins e Dayan, 1992) é um dos algoritmos de Aprendizado por Reforço mais conhecidos. O método (Algoritmo 1) se baseia na atualização da matriz de aprendizado Q, a partir da Equação 1.

$$Q_{t+1} = Q_t(s,a) + \alpha[r(s,a) + \gamma max'_a Q(s') - Q_t(s,a)], \quad (1)$$

em que: Q é a matriz de aprendizado, $r(s,a)$ é o reforço para a execução da ação a no estado s, α é a taxa de aprendizado e γ é o fator de desconto.

A matriz Q possui dimensões $n_s \times n_a$, em que n_s é o número de estados do modelo e n_a é o número de ações que o agente pode executar. Assim, para cada tomada de decisão (ação) em um determinado estado do ambiente, é retornado ao agente um sinal $r(s,a)$. Dessa forma, em cada instante de tempo t, a matriz é atualizada em único par estado $\times$ ação.

Para a atualização de Q também é necessário identificar o novo estado do ambiente (s' ou s_{t+1}). Já que, o termo $max'_a Q(s')$ é o máximo valor na linha relativa ao novo estado s' na matriz.

Já o algoritmo SARSA (Sutton e Barto, 1998) é uma modificação do *Q-learning*. O SARSA (Algoritmo 2) não adota a maximização das ações do *Q-learning*, assim a matriz de aprendizado é atualizada como na Equação 2:

$$Q_{t+1} = Q_t(s,a) + \alpha[r(s,a) + \gamma Q_t(s',a') - Q_t(s,a)], \quad (2)$$

em que: $s' = s_{t+1}$ e $a' = a_{t+1}$.

Nos Algoritmos 1 e 2, a política denominada $\epsilon - gulosa$ (ou, quase-guloso) é responsável pelo controle entre gula e aleatoriedade na seleção das ações (Sutton e Barto, 1998).

Definir os parâmetros: α, γ e ε
Para cada ***s,a*** inicialize ***Q(s,a)***=0
Observe o estado ***s***
repita
  Selecione a ação ***a*** usando a política ε-gulosa
  Execute a ação ***a***
  Receba a recompensa imediata ***r(s,a)***
  Observe o novo estado ***s'***
  Atualize Q(s,a) com a Eq. (1)
  s = ***s'***
até *o critério de parada ser satisfeito*;

Algoritmo 1: *Q-learning.*

Definir os parâmetros: α, γ e ε
Para cada ***s,a*** inicialize ***Q(s,a)***=0
Observe o estado ***s***
Selecione a ação ***a*** usando a política ε-gulosa
repita
  Execute a ação ***a***
  Receba a recompensa imediata ***r(s,a)***
  Observe o novo estado ***s'***
  Selecione a nova ação ***a'*** usando a política ε-gulosa
  Atualize Q(s,a) com a Eq. (2)
  s = ***s'***
  a = ***a'***
até *o critério de parada ser satisfeito*;

Algoritmo 2: SARSA.

3 Problema do Caixeiro Viajante com Reabastecimento

O Problema do Caixeiro Viajante (PCV) visa encontrar o menor caminho entre um conjunto de localidades, passando por cada cidade uma única vez. Já o Problema do Caixeiro Viajante com Reabastecimento (PCVR) é uma variação do PCV, e tem por objetivo encontrar a rota para minimizar o custo com combustível (Khuller et al., 2007).

3.1 Instâncias Minas24 e Minas30

Neste trabalho, são propostas duas novas instâncias para o PCVR: Minas24 e Minas30. Cada uma delas envolve um conjunto de cidades do Estado de Minas Gerais (BRA). As instâncias propostas levam em consideração as distâncias entre cada localidade e o custo médio de combustível em cada uma delas. As distâncias entre cada cidade foram calculadas a partir das posições em latitude e longitude. Já os preços dos combustíveis foram pesquisados em Novembro/2015 no site da Agência Nacional do Petróleo. Em seguida, as cidades que compõem cada uma das instâncias:

- Minas24: Araguari, Belo Horizonte, Betim,

Campo Belo, Contagem, Formiga, Governador Valadares, Guaxupé, Itabira, Ituiutaba, Juiz de Fora, Monte Carmelo, Montes Claros, Oliveira, Patos de Minas, Poços de Caldas, Pouso Alegre, Sete Lagoas, Teófilo Otoni, Três Corações, Uberaba, Uberlândia, Unaí e Varginha.

- Minas30: Localidades de Minas24 mais Araxá, Barbacena, Divinópolis, Ipatinga, Lavras e Passos.

3.2 Características do Veículo

O veículo fictício considerado neste trabalho busca representar um automóvel de carga de pequeno porte, estilo caminhonete (Ex. S10), com consumo de gasolina. Foi cosiderado uma taxa de consumo constante de 10 km/litro e capacidade máxima do tanque de 80 litros.

4 Modelagem do Sistema Híbrido

4.1 Modelagem do Sistema de Aprendizado por Reforço

A metodologia adotada é baseada em Ottoni et al. (2015) e Lima Júnior et al. (2010):

1. Definição dos estados do ambiente: Os estados foram definidos como todas as localidades em que o caixeiro viajante (agente) deve acessar (Ottoni et al., 2015).
2. Definição do conjunto finito de ações que o agente pode realizar: Cada ação foi definida como sendo intenção de ir para outra localidade (estado) do problema (Ottoni et al., 2015).
3. Definição dos valores dos reforços. Neste trabalho, os reforços visam oferecer um retorno ao caixeiro viajante de acordo com a distância entre as localidades, o nível de combustível e o valor de combustível nas cidades de reabastecimento. Assim, os reforços são maiores para cidades mais próximas ($d \leq 250$), e mais negativos para cidades mais distantes ($250 < d \leq 850$ e $d > 850$). Além disso, para as cidades mais próximas, os reforços variam de acordo com o nível do combustível na cidade de partida (A) e o preço do combustível na cidade de chegada (B).

   Foram definidos dois métodos de reforço: sem fuzzy e com fuzzy. O primeiro método de reforço é apresentado na Tabela 1, em que:

   - Dist. (d): é a distância entre duas localidades A e B. Sabendo que o caixeiro vai de A à B.
   - Nível (n): é o nível de combustível no tanque do automóvel, na cidade de partida A.
   - Preço (p): é o preço do combustível na localidade de chegada B.

Tabela 1: Valores dos reforços.

Dist. (d)	**Nível** (n)	**Preço** (p)	**R**
$d \leq 250$	$n \geq 0,5$	$3{,}40 \leq p < 4,00$	1300
$d \leq 250$	$n < 0{,}5$	$3{,}40 \leq p < 3,51$	2000
$d \leq 250$	$n < 0{,}5$	$3{,}51 \leq p < 3,63$	1300
$d \leq 250$	$n < 0{,}5$	$3{,}63 \leq p < 3,76$	650
$d \leq 250$	$n < 0{,}5$	$3{,}76 \leq p < 3,90$	0
$d \leq 250$	$n < 0{,}5$	$3{,}90 \leq p < 4,00$	-650
$250 < d \leq 850$	$0 \leq n \leq 1$	$3{,}40 \leq p < 4,00$	-1300
$d > 850$	$0 \leq n \leq 1$	$3{,}40 \leq p < 4,00$	-2000

Já a próxima seção descreve o método de modelagem da função recompensa com lógica fuzzy.

4.2 Modelagem da Função de Recompensa com Lógica Fuzzy

Nesta seção, será apresentada a metodologia adotada para a modelagem da função de recompensa com lógica fuzzy.

Em seguida, são apresentadas as oito regras definidas no modelo:

1. Se a (Distância é Perto) e o (Nível do Tanque é Normal) então o (Reforço é Alto);
2. Se a (Distância é Perto) e o (Nível do Tanque é Baixo) e o (Preço do Combustível é Muito Baixo) então o (Reforço é Muito Alto);
3. Se a (Distância é Perto) e o (Nível do Tanque é Baixo) e o (Preço do Combustível é Baixo) então o (Reforço é Alto);
4. Se a (Distância é Perto) e o (Nível do Tanque é Baixo) e o (Preço do Combustível é Médio) então o (Reforço é Meio Alto);
5. Se a (Distância é Perto) e o (Nível do Tanque é Baixo) e o (Preço do Combustível é Alto) então o (Reforço é Médio);
6. Se a (Distância é Perto) e o (Nível do Tanque é Baixo) e o (Preço do Combustível é Muito Alto) então o (Reforço é Meio Baixo);
7. Se a (Distância é Média) então o (Reforço é Baixo);
8. Se a (Distância é Grande) então o (Reforço é Muito Baixo).

Em seguida, foram definidas as funções de pertinência para as três variáveis de entrada (ver Figuras 1, 2 e 3) e para a variável de saída (ver Figura 4) do modelo. Já a Figura 5 apresenta os resultados da modelagem fuzzy para a função de recompensa.

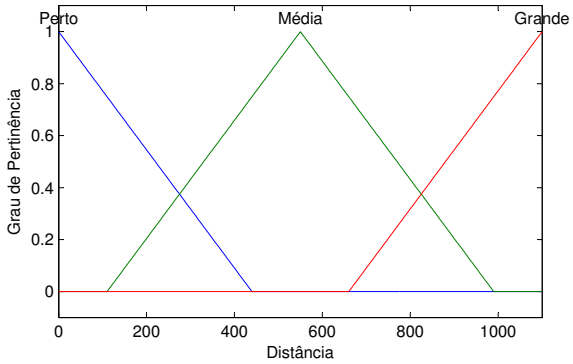

Figura 1: Funções de pertinência da variável de entrada Distância.

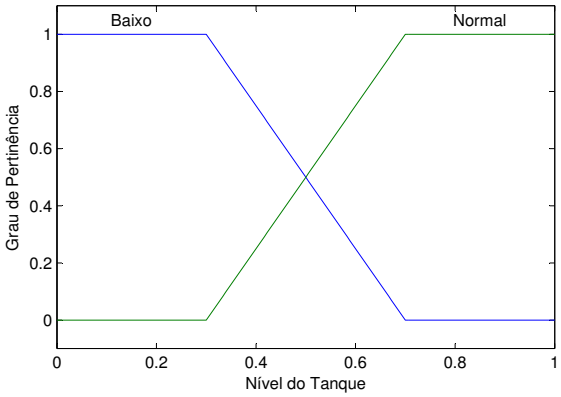

Figura 2: Funções de pertinência da variável de entrada Nível do Tanque.

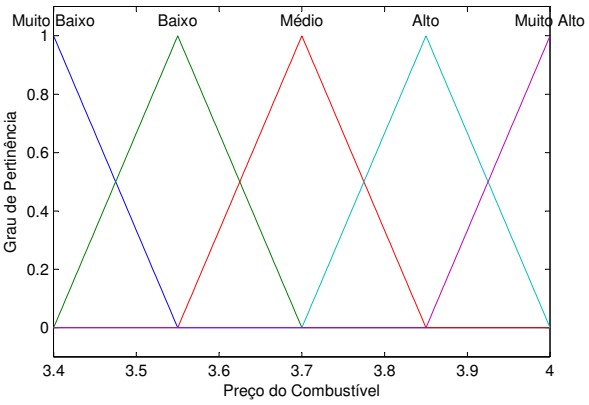

Figura 3: Funções de pertinência da variável de entrada Preço do Combustível.

5 Experimentos Realizados

Para a solução das instâncias Minas24 e Minas30 foram analisados quatro métodos:

- *Q-learning* com o primeiro método de reforço;

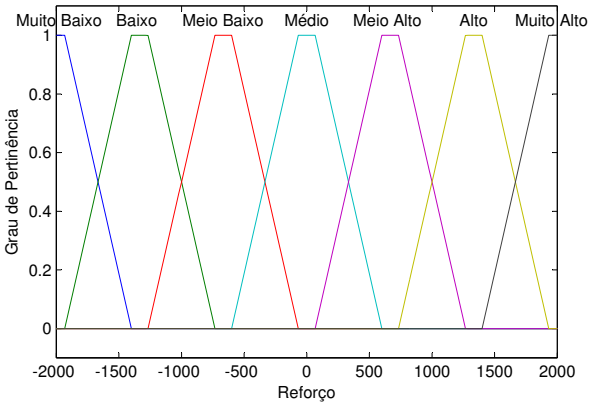

Figura 4: Funções de pertinência da variável de saída Reforço.

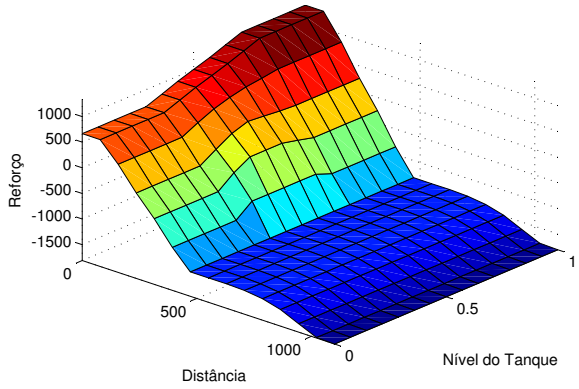

Figura 5: Superfície 3D que retrata o Reforço em função da Distância (km) e o Preço do Combustível (R$).

- Q-Fuzzy: *Q-learning* com função de recompensa fuzzy;
- SARSA com o primeiro método de reforço;
- S-Fuzzy: SARSA com a função de recompensa fuzzy.

Os parâmetros do aprendizado por reforço foram definidos de acordo com os experimentos e análises de Ottoni et al. (2015):

- taxa de aprendizado decaindo $\alpha_n(s,a) = \frac{1}{1+visitas_n(s,a)}$;
- fator de desconto (γ) = 0,01;
- política $\epsilon - greedy$ = 0,01;

Vale ressaltar que, cada simulação foi padronizada em três épocas (repetições) com 5000 episódios. Sendo que, cada episódio teve como resposta a distância total percorrida pelo agente na

rota da instância e o gasto em reais em reabastecimento na rota. Os experimentos foram realizados no *software* $MATLAB$®.

6 Análise dos Resultados

Em seguida, a análise de desempenho do sistema de aprendizado por reforço para cada uma das duas instâncias estudadas: Minas24 e Minas30.

6.1 Resultados com 24 localidades - Minas24

Para a instância Minas24, os melhores resultados calculados foram R$1085,2 e 3583,7 km. Esses valores foram encontrados quando adotado o método S-Fuzzy. A Tabela 2 apresenta os resultados mínimos encontrados de acordo com o algoritmo.

Tabela 2: Melhores resultados para a instância Minas24.

Algoritmo	**Gasto (R$)**	**Distância (km)**
Q-learning	1592,5	4819,6
Q-Fuzzy	1250,2	3951,4
SARSA	1465,0	4488,0
S-Fuzzy	1085,2	3583,7

6.2 Resultados com 30 localidades - Minas30

Para a instância Minas30, os melhores resultados calculados foram R$1265,30 e 3935,6 km. Esses valores também foram encontrados quando adotado o método S-Fuzzy. A Tabela 3 apresenta os resultados mínimos encontrados de acordo com o algoritmo.

Tabela 3: Melhores resultados para a instância Minas30.

Algoritmo	**Gasto (R$)**	**Distância (km)**
Q-learning	1917,6	5748,2
Q-Fuzzy	1445,4	4349,0
SARSA	1897,8	5639,4
S-Fuzzy	1265,3	3935,6

Apesar dos resultados dos métodos SARSA e S-Fuzzy serem semelhantes no início do aprendizado (Figura 6), a Figura 7 revela que ao longo de uma sequência maior de episódios, o método que adotou a função de recompensa fuzzy alcançou desempenho superior. Já as Figuras 8 e 9, mostram o comportamento do AR, a partir dos caminhos encontrados para a instância Minas30, para a melhor solução com o algoritmo SARSA (Figura 8), e melhor solução para o método S-Fuzzy (Figura 9).

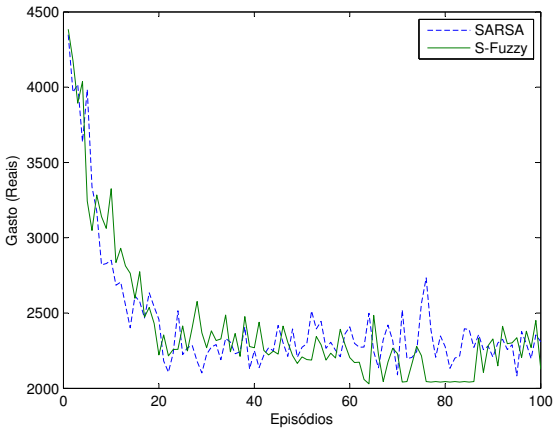

Figura 6: Média de Gasto em Reais versus o episódio de aprendizado, para a instância Minas30 e algoritmos algoritmos SARSA e S-Fuzzy. Visualização dos 100 episódios iniciais.

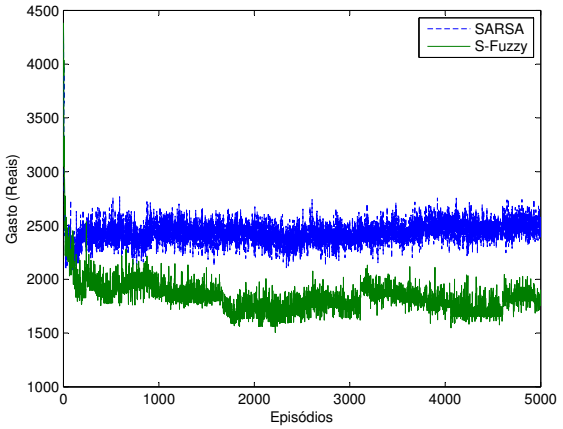

Figura 7: Média de Gasto em Reais versus o episódio de aprendizado, para a instância Minas30 e algoritmos SARSA e S-Fuzzy.

7 Conclusão

Este trabalho teve como objetivo principal desenvolver um modelo híbrido de Aprendizado por Reforço e Lógica Fuzzy para aplicação no Problema do Caixeiro Viajante com Reabastecimento.

Os métodos com recompensa fuzzy, Q-Fuzzy e S-Fuzzy, apresentaram os melhores desempenhos. Além disso, vale ressaltar a superioridade dos algoritmos SARSA sobre o *Q-learning* nos resultados das duas instâncias.

Em trabalhos futuros, espera-se analisar instâncias maiores para esse problema, aumentando o nível de dificuldade de resolução. Além disso, serão analisados outras formas de recompensa para o PCVR.

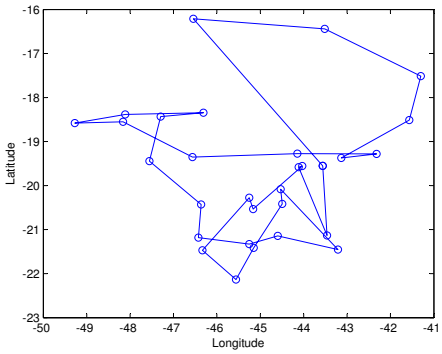

Figura 8: Caminho definido pelo algoritmo SARSA para a instância Minas30. Distância percorrida igual a 5639,4 km e gasto de R$1897,8.

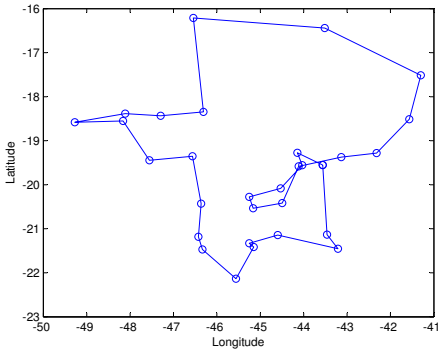

Figura 9: Caminho definido pelo algoritmo S-Fuzzy para a instância Minas30. Distância percorrida igual a 3935,6 km e gasto de R$1265,3.

Agradecimentos

Agradecemos à CAPES, CNPq, FAPEMIG e UFSJ pelo apoio.

Referências

Bianchi, R. A. C. (2004). *Uso de Heurística para a aceleração do aprendizado por reforço.*, PhD thesis, Escola Politécnica da Universidade de São Paulo.

Bonarini, A., Lazaric, A., Montrone, F. e Restelli, M. (2009). Reinforcement distribution in fuzzy q-learning, *Fuzzy Sets and Systems* **160**(10): 1420 – 1443.

Derhami, V., Majd, V. J. e Ahmadabadi, M. N. (2010). Exploration and exploitation balance management in fuzzy reinforcement learning, *Fuzzy Sets and Systems* **161**(4): 578 – 595.

Faria, G. e Romero, R. A. F. (2002). Navegação de robôs móveis utilizando aprendizado por reforço e lógica fuzzy, *Revista Controle & Automação* **13**(3): 219–230.

Khuller, S., Malekian, A. e Mestre, J. (2007). *Algorithms – ESA 2007: 15th Annual European Symposium, Eilat, Israel, October 8-10, 2007. Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, chapter To Fill or Not to Fill: The Gas Station Problem, pp. 534–545.

Lima Júnior, F. C., Neto, A. D. D. e Melo, J. D. (2010). *Hybrid Metaheuristics Using Reinforcement Learning Applied to Salesman Traveling Problem, Traveling Salesman Problem, Theory and Applications, Prof. Donald Davendra (Ed.)*, InTech.

Lin, S. H., Gertsch, N. e Russell, J. (2007). A linear-time algorithm for finding optimal vehicle refueling policies., *Operations Research Letters* **35**(3): 290–296.

Ottoni, A. L. C., Nepomuceno, E. G., Cordeiro, L. T., Lamperti, R. D. e Oliveira, M. S. (2015). Análise do desempenho do aprendizado por reforço na solução do problema do caixeiro viajante, *XII SBAI - Simpósio Brasileiro de Automação Inteligente.* pp. 43–48.

Rodrigues Junior, A. D. (2011). *Um modelo de otimização da política de reabastecimento para transportadores rodoviários de carga*, Master's thesis, Programa de Pós-Graduação em Engenharia Civil da UFES.

Rodrigues Junior, A. D. e Cruz, M. M. C. (2013). A generic decision model of refueling policies: a case study of a brazilian motor carrier, *Journal of Transport Literature* **7**(4): 8–22.

Silva, H. L. F. e Cruz, M. M. C. (2014). Uma revisão de literatura sobre problemas de reabastecimento de veículos transportadores de cargas, *XXVIII Congresso de Pesquisa e Ensino em Transportes.* pp. 1–9.

Sutton, R. e Barto, A. (1998). *Reinforcement Learning: An Introduction*, 1st edn, Cambridge, MA: MIT Press.

Suzuki, Y. (2008). A generic model of motor-carrier fuel optimization, *Naval Research Logistics* **55**(8): 737–746.

Suzuki, Y. (2012). A decision support system of vehicle routing and refueling for motor carriers with time-sensitive demands, *Decision Support Systems* **54**(1): 758–767.

Watkins, C. J. e Dayan, P. (1992). Technical note q-learning, *Machine Learning* .

Zhou, C. e Meng, Q. (2003). Dynamic balance of a biped robot using fuzzy reinforcement learning agents, *Fuzzy Sets and Systems* **134**(1): 169 – 187.