Intelligent Agent Based Traffic Signal Control on Isolated Intersections

Daniela Koltovska, Kristi Bombol

St. Kliment Ohridski University, Faculty of Technical Sciences, Department for Traffic and Transport, POB 99 7 000 Bitola, Republic of Macedonia

Abstract – **The purpose of this paper is to develop an adaptive signal control strategy on isolated urban intersections. An innovative approach to defining the set of states dependent on the actual and primarily observed parameters has been introduced. The Q–learning algorithm has been applied. The developed self-learning adaptive signal strategy has been tested on a real intersection. The intelligent agent results have been compared to those in cases of fixed-time and actuated control. Regarding the average total delay, the total number of stops and the total throughput, the best results have been obtained for unknown traffic demand and over-capacity.**

Keywords – **adaptive control, isolated intersections, artificial intelligence, intelligent agent, *Q*-learning.**

1. Introduction

1.1. Research background

Nowadays there are complex systems of various generations utilized for urban traffic control. Using different approaches to the urban traffic signal control problem, each generation provides improved functionality and flexibility than its predecessors. Over the past 30 years of development, improvement and practical application, the existing systems have displayed several weaknesses, such as [1], [3]:

1. Shortcomings in the accuracy and range of application of traffic prediction models
2. Limited ability to effectively deal with a complete range of traffic conditions, including oversaturated conditions
3. Limited ability to adapt to a changing environment (beyond perceived variations in traffic demand)
4. Limited ability to deal with traffic conditions spatially or temporally removed from a local control decision yet affecting it or affected by it
5. Present adaptive systems are related to the corridor traffic control or to that on the network level rather than to isolated intersection. Thus, the solution for an isolated intersection is below optimal or inapplicable. The problem of an isolated intersection is still very acute due to the fact that the number of light control isolated intersections surpasses 50% in a large number of countries throughout the world (particularly in Europe) [10].
6. This is why Gartner emphasizes the need of new concepts development instead of extension of current ones. This implies introduction of 4th and 5th generation which are embodying levels of "intelligence" higher the ones achieved to date [3], [5].

The 4-LC system focuses on intelligence such as dynamic traffic assignment capability for proactive control and traffic event responsive. The 5-LC seems to be the most intriguing one as it is the *"super level"*of systems incorporating control strategies for self-learning on the basis of the experience developed with artificial intelligence (AI) techniques [3].

Reinforcement learning (RL) is a technique well known in AI and machine learning (ML) communities. It has the potential to provide: a) the defined functionality of 5-LC, b) solutions to some of the major problems with adaptive systems.

In order to evaluate the effects of the *Q*-learning algorithm implementation, Abdulhai et al have tested this algorithm for an isolated intersection traffic control [2]. The learning agent has been applied on the intersection with through movements only. The queue length and the time elapsed from the last phase change have been used as input parameters. There is no information on how the modeling has been performed. The strategy performance has been compared with the fixed time control. They have continued their research by applying *Q*-learning and multi agents' technique. The advantage of the multi-agents' technique is that its control distribution makes the system robust rather than centralized (even in situations of communication problems). This paper outlines an encompassing and clear review of the RL and *Q*-learning concept as well as the possibilities arising for their application in the development of the fifth generation of control strategies.

The effectiveness that $Q(\lambda)$ has in adaptive traffic control has been examined further on [7]. Delays are taken as $Q(\lambda)$ states, whereas green duration as the action. The $Q(\lambda)$ learning performances are being compared with the fixed time control. The results display small time delays in variable traffic conditions.

The research referred to the above confirms the potential advantages of the RL. They open up new horizons for the development of innovative self-learning strategies [4].

1. In the examinations so far, the queue lengths or the delays or the number of stops etc., have been taken as input parameters, which is very difficult to be measured in real time. In describing the states of environment for the agent, the following three variables have been considered: the phase, the time gap, and the inductive loop detector occupancy. The reward function has been suggested in a new manner – maximization of throughput, which indirectly influences the reduction of delays (as compared to previous research activities).
2. The biggest share of research exploration activities have been conducted on hypothetical intersections.
3. The analyzed strategies do not have the *feature of self-learning* and *self-adapting to the changes in the environment.*
4. The process of developing adaptive control strategies which apply the artificial intelligence techniques, is by no means simple, particularly not for traffic engineering researchers.

It can be concluded that the research is to be continued in the direction of designing adaptive control strategies which do not require *traffic prediction model, environment model* or *developing strategies* in terms of *self-learning* and *self-adaptation in direct interaction with the environment* [4]. Applying the *AI techniques* and *algorithms in the area of ML* opens up opportunities for developing adaptive control strategies.

2. Theoretical Background

"Learning to act in ways that are rewarded is a sign of intelligence. It is, for e.g., natural to train a dog by rewarding it when it responds appropriately to commands. That animals can learn to obtain rewards and to avoid punishments is generally accepted". (Watkins, 1989) [9].

The above stated citation expresses the essence of RL, that is: "Reinforcement learning is learning what to do-how to map situations to actions-so as to maximize a numerical reward signal". (Sutton et al, 1998) [8].

Inspired by behaviourist psychology, RL is tightly related to Psychology, Neurological Sciences, Artificial Neural Networks, Control Theory and Operational Research, AI planning methods; the links with the last two fields being much stronger (Figure 1). There are various problems that can be solved by applying RL. Requiring no supervision when learning, RL agents show best when used in complex problems with no apparent and easy programme solution. (E.g.: non-linear systems control, computer games, robotics etc.).

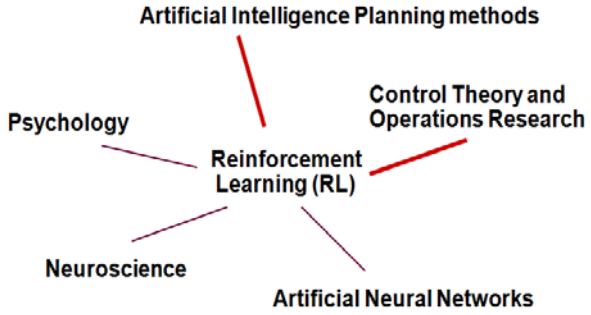

Figure 1. Reinforcement learning and relationships to other fields

RL is one of the basic techniques of the intelligent agent (IA) technology**.** The learner or decision maker is named agent, and everything it interacts with is named environment. The agent has a set of sensors to observe the state of the environment, and to perform a set of actions in order to change the state of the environment. The most important characteristics of the agent are trial and error search, and delayed reward [8].

The learner or an autonomous agent that senses its environment or acts in it can learn through trials to select the optimal action or actions which lead to the highest reward.

For a more accurate presentation of the interaction we here assume that the agent and the environment communicate in each sequence of discrete time steps: t=0,1,2,… In each time step, t, the agent receives some representation of the state of the environment, $s_t \in S$, where S is the set of possible states. In accordance with that, action $a_t \in A(s_t)$ is chosen, where $A(s_t)$ is a set of actions which are available in the state s_t. One step later, as a consequence of its action, the agent gets a numerical reward, $r_{t+1} \in R$ and finds itself in a new state, S_{t+1}. The agent obtains a reward or a penalty in order to induce the desirability of the final state. Figure 2 shows the agent-environment interaction.

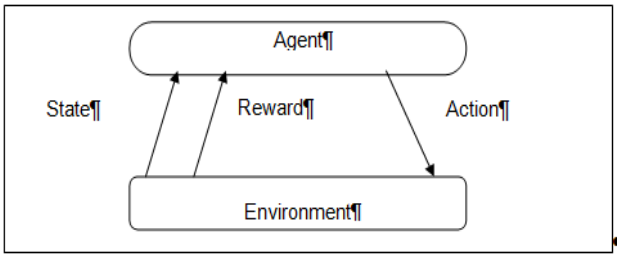

Figure 2. Agent-Environment Interaction

The transition from one to another state is shown as:

$$s_0 \xrightarrow[r_0]{a_0} s_1 \xrightarrow[r_1]{a_1} s_2 \xrightarrow[r_2]{a_2} \quad (1)$$

Where S_i is the state in the time step i, a_i is the possible action available in each state in the time step i, r_i is the reward which the agent receives in the time step i for taking action a_i.

In addition to the agent and its environment, other elements that can be distinguished are *policy*, *reward function*, *value function*, *model of environment.*

One of the most significant achievements in RL was the development of Temporal Differences Off-Policy Algorithm known as *Q*-learning. This algorithm was developed by Watkins in 1989 [8]. It has been the most studied one both theoretically and practically.

The control strategy developed with this research is performed by an agent. In order to embody the learning feature into the agent, the RL technique and *Q*-learning algorithm have been applied.

3. Research Methodology

The process of developing adaptive control strategy for an isolated intersection is composed of three steps (Figure 3) [4], [6]:

Step 1: **Development of model**
Step 2: **Design and development of the Intelligent Agent (IA)**
Step 3: **Strategy Testing and Evaluation**

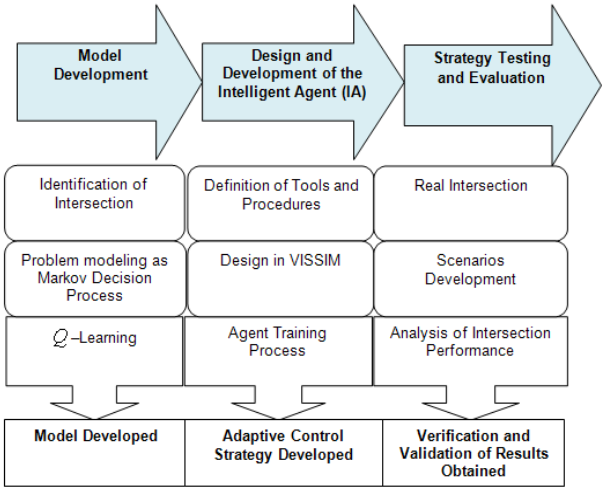

Figure 3. Methodology of adaptive control strategy development process

Step 1: **Model Development**

When Markov decision process exists, the process of RL can be applied. In this context, to have the RL agent learn the control policy (to take decisions for changing the traffic signal states), it is necessary to determine the *set of states.*

Defining the set of states S

The selection of the variables to describe the traffic process greatly varies. Within the research, *phase*, *gap*, *occupancy* are applied.

The set of states S has been defines as $S = \{(\phi, g, Occ);\ \phi \in \{1,2\},\ g \in \{YES, NO\},\ Occ \in \{0,1\}\}$, where ϕ is the signal phase within a signal cycle of C = 90 seconds); when ϕ = 1 it is a green phase, when ϕ = 2, the phase is red. Green time t_g, within a single signal cycle C of 90 seconds falls within the interval of 24 to 78 seconds, i.e. $t_g \in [24,78]$. Red time t_r within a single signal cycle C of 90 seconds falls within the interval of 12 to 66 seconds, i.e. $t_r \in [12,66]$; g is a binary variable receiving the values $\{YES, NO\}$, where the value of NO denotes that there are no vehicles (signal received from the inductive loop), YES represents the opposite; Occ is a binary variable, where the value of 0, denotes that there are no present vehicles from the conflict flow (red light), and the value of 1, denotes the opposite.

Defining the set of actions A

Based on the information related to the detected state, the control agent takes up action. For each state, the agent can only take up two actions: action value of 1, which means the state remains the same (green time extension), or action value of 0, which means change of the signal state.

Defining the set of rewards R

The rewarding function is the second key element for the agent. The reward is a function that depends on the system's state and the action taken. The reward takes values from the set of natural numbers, i.e. it is defined as mapping $R: A \times S \rightarrow$.

The rewarding function goal is maximization of the total throughput. For that purpose, the following set of rewards was defined:

1. **Reward Function** – the total throughput
2. **Immediate reward** – the number of vehicles passing at green light in the **previous time interval** (the length of this interval is 90 seconds)
3. **Discounted reward** – total number of vehicles in a 3600 second cycle (peak hour for which the testing is made)

The action is taken at a shorter interval for a given time step. The vehicles are counted per one 90

second signal cycle. An action is taken per second – over the green time duration, in which case the step takes 3 seconds.

Q-learning provides the agent with an opportunity to learn the control policy: to select actions for changing the signals so that it can bring about a maximum throughput as well as reduction of delays at the intersection.

For non-deterministic environment the *Q*-function has been redefined as an expected value $\hat{Q}_n(s,a)$ from a previous defined value for deterministic case.

By applying the learning rule

$$\hat{Q}_n(s,a) \longleftarrow (1-\alpha_n)\hat{Q}_{n-1}(s,a) + \alpha_n[r + \gamma \max_{a'} \hat{Q}_{n-1}(s',a')] \quad (2)$$

Whereas the learning rate is

$$\alpha_n = \frac{1}{1+visits_n(s,a)} \quad (3)$$

$\hat{Q}_n$ still converges to *Q** whereas *Q** is the optimal action value function, $\hat{Q}_n(s,a)$ is the expected value of the previous defined value for deterministic function case for action *a* and state *s* and $\hat{Q}_{n-1}(s',a')$ is the expected value of the previous defined value for the new action *a'* in the next state *s'* [8]. The parameter γ is the discount rate in the range of $0 \leq \gamma \leq 1$, α_n is the learning rate, *(s,a)* is the updated state and action during *n* iterations, and $visits_n$ *(s,a)* is the total number of visits for this pair of state-action until the n^{th} iteration.

This research uses the look-up table to describe the *Q*-function (Figure 4). The look-up table is a matrix (4 X 4) that is being created in the course of learning upon agent's receiving rewards (that is to say, learning). The rows in the *Q*-matrix present the current state of the agent, and the columns present the actions directing to the next state. Initially, the values of the actions are adjusted to display zero, though they may be adjusted as random values.

Q-matrix is the brain of our agent and it stores the memory of what the agent has learnt via numerous trials. This approach of describing the *Q*-function in the look – up table is simple to use. However, in the case of large space of states, difficulties may arise because of the use of a huge space for memory.

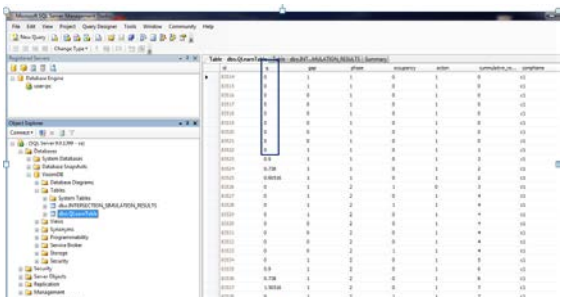

Figure 4. Look-up Table of Q values

The values of the parameters α and γ are displayed in Table 1.

Table 1. Parameters α and γ values applied in the research.

Parameter	Parameter values
α	- In the first third of the process of learning, the parameter value is 0.9; in the next third, the iterations are run with parameter value of 0.6; and in the final third, the total of iterations are run with parameter value of 0.3. There is also an option for a gradient change of α, for e.g., at each 100th iteration lowering the value of α to 0.1 . - *Comment*: The point of changing α is changing the balance of the exploration/exploitation relationship that the algorithm is to be determined whereby in the beginning the algorithm is set to do more search for the solution, and later it is set to optimize the solution it has found.
γ	- The parameter is responsible for the reward transfer. It determines the influence of the future rewards over the agent's behavior. - Value of γ is 80 .

Each action, derived by the agent, influences the environment; upon the completion of the action, the environment is at a new state. For each action taken, the agent is rewarded and the reward defines the extent to which the action was good or bad. The rewarding helps the agent to learn what to do and to act in a *more intelligent manner*.

Step 2: **Design and Development of the Intelligent Agent (IA)**

a) Defining the tools and procedures

In the design of the intelligent agent, the following tools have been used:

- VISSIM5.4-0.3(**V**erkehr**I**n **S**tädten**SIM**ulationsmodell; "Traffic in Cities - simulation model").
- VISSIM **COM** (**C**OMPONENT **O**BJECT **M**ODEL).
- Microsoft SQL (MSSQL) .

b) Design by VISSIM simulator

To learn the control strategy, the RL agent requires a simulated traffic system environment. The simulation platform that is being used is VISSIM. The traffic demand has been created via the simulator's graphical interface. The number of vehicles is entered for every link at intervals of 15 minutes per peak hour (known/unknown demand for intelligent agent). Vehicle arrivals are described by the Poisson distribution.

As the strategy applies for the peak hour, the simulation period is 1 hour (3600 seconds). To express the stochastic variations of traffic flows as realistically as possible, the parameter used to initialize a random number generator is applied (*Random Seed*).

From the user side, the number, the position and the detector dimension are defined by applying the simulator's graphic interface. Each detector is connected with a corresponding signal and a corresponding phase.

The program for communication among VISSIM simulator, the database and the RL algorithm is developed using the C Sharp (C#) program language (Figure 5).

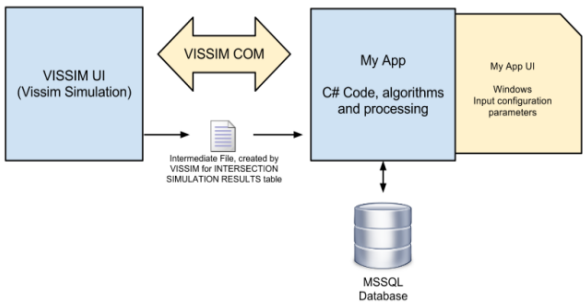

Figure 5. The process of communicating and interaction among the main elements

Figure 6 shows the sequence diagram.

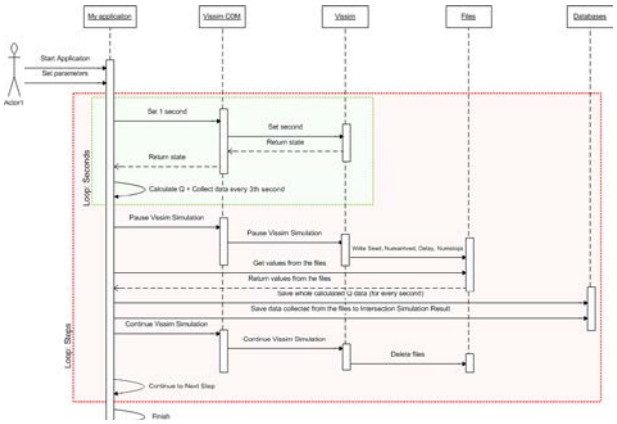

Figure 6. Sequence diagram

The agent is being trained in simulation conditions. However, after being applied in the field, the agent can continue to learn, starting with the last *Q*–values obtained in the training process.

After sufficient number of iterations and convergence of *Q*–values, the training phase is completed. The next phase is testing and evaluation of the adaptive control strategy.

4. Microsimulation Based Strategy Testing and Evaluation of Adaptive Signal Control Strategy

The strategy testing is performed on a real four-leg intersection located within the central area of Bitola, with real traffic data. Figure 7 depicts the intersection and the communication with the RL intelligent agent.

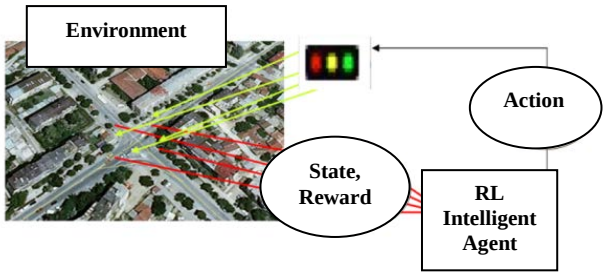

Figure 7. Description of intersection and communicating with the RL agent

Delay, *throughput* and *number of stops* are analyzed as strategy efficiency measures.

The results obtained from the learning IA are compared to the ones obtained through simulations in cases of fixed time and actuated control. The fixed time control is selected as a base case and all the other results are estimated in relation to it.

The testing is performed after three hundred of iterations with various values regarding states and after the convergence of *Q*–values. When testing, the selected action is the one with maximum *Q* value and the one that will provide optimum control action in all of the agent states.

Depending on the traffic flow conditions, and

whether the traffic demand is known or unknown to the agent, the testing is performed in two phases. During the first phase, the testing is performed for uncongested traffic conditions with known and unknown demand. During the second phase, the testing is performed for congested traffic conditions with known and unknown demand.

Figure 8 shows the comparison of percentage of efficiency measure improvements for all phases of testing in case of applying fixed-time to that of adaptive control in conditions of traffic non-congestion and congestion, with known and unknown demand.

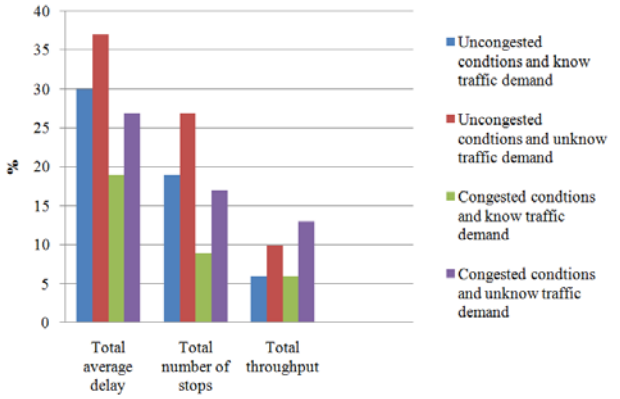

Figure 8. Comparison of percentage of efficiency measure improvements (fixed-time / adaptive control)

Figure 8 shows that adaptive strategy gives best results in cases of:

- Total average delay of vehicles (37%) and the total number of stops (27%) in uncongested traffic conditions for unknown traffic demand
- Total throughput (13%) in congested traffic conditions for unknown traffic demand

Figure 9 displays the comparison of improvements for all testing stages in cases of applied fixed-time as opposed to actuated control, in both uncongested and congested traffic conditions, for known and unknown traffic demand.

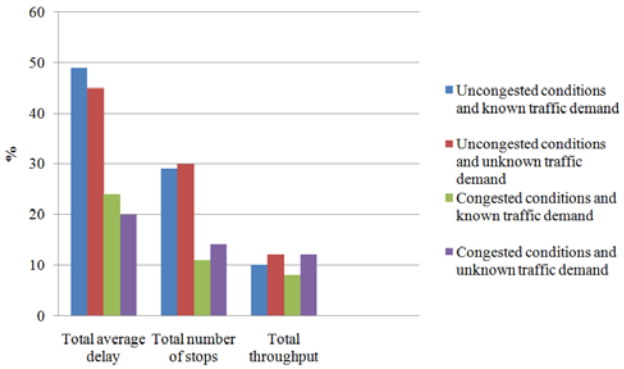

Figure 9. Comparison of percentage of efficiency measure improvements (fixed-time/actuated control)

What Figure 9 shows is that actuated control renders best results with total average delay in uncongested traffic conditions and known traffic demand (49%), with total number of stops in uncongested and unknown traffic demand (30%) and with total throughput in uncongested traffic conditions and unknown traffic demand (12%).

Overall, the following has been observed:

- The adaptive strategy gives best results with total average delay (37%) and with a total number of stops (27%), *in uncongested traffic conditions for unknown traffic demand*
- The adaptive strategy gives best results with the total throughput (13%) *in congested traffic conditions for unknown traffic demand*
- The actuated control gives best results with total average delay *in uncongested traffic conditions for known traffic demand* (49%)
- The actuated control gives best results with total number of stops *in uncongested traffic conditions for unknown traffic demand* (30%)
- The actuated control gives best results with total throughput *in uncongested traffic conditions for unknown traffic demand* (12%)

As regards all efficiency measures (total average delay, total number of stops and total throughput), best output results are obtained with the newly designed adaptive control strategy in cases of unknown traffic demand in congested traffic conditions of over-capacity.

Summary

This scientific research refers to a new extension of the well-known approaches by applying *Q*-learning to the development of traffic signal control strategies. An innovative approach to defining the set of states dependent on the actual and primarily observed parameters has been introduced. The *Q*–learning algorithm has been applied in the development of a self-learning adaptive traffic signal control on isolated intersection.

The developed self-learning adaptive signal strategy has been tested on a real four-leg urban intersection. The intelligent agent results have been compared to those in cases of fixed-time (base case) and actuated control. Depending on a) the traffic flow conditions, and b) the known and unknown demand, the testing has been performed for non-congestion and over-capacity. Regarding the average total delay, the total number of stops and the total throughput, the best results have been obtained for unknown traffic demand and over-capacity.

Having in mind the testing results it can be deduced that the *newly designed adaptive control strategy is appropriate for controlling the traffic at isolated urban intersections.* In favour of this speaks the comparison of results obtained in every of the testing stages and scenarios.

Regarding all three efficiency measures (total average delay, total number of stops and total throughput), there are evident improvements that are achieved by means of the newly designed adaptive control strategy for *unknown traffic demand in over-capacity congested traffic conditions.*

Based upon the above, it can be concluded that the newly designed strategy for isolated intersections in urban areas is well adapted to the traffic flow conditions (feature of adaptability) and depends on the real-time traffic demand (responds to the demand).

5.1 Limitations and Further Work

Albeit the developed adaptive traffic signal control on isolated intersection has rendered encouraging results, several limitations that were in the way have to be mentioned:

1. Pedestrians were not taken into account
2. The queue length was taken into consideration within the frames of minimum green time duration. But, in conditions of queue increasing to an extent that would negatively influence the operations upstream, a precise model for detecting the queue length is necessary

Nevertheless, this research represents a sound basis for further exploration in the area of control strategies, which are self-learning from the interaction with the environment and self-adaptive to the real-time traffic flow conditions. The following directions for future research activities are recommended:

1. Developing a scenario that involves two inductive loop detectors per each lane. In this case, the goal function would be minimization of the queue length, whereas the reward function would be defined as penalty. This would mean that if the queue is getting longer, the agent would be punished
2. Applying the *transfer learning* approach to neighbouring intersections
3. Exploring the strategy efficiency after its application in the field – testing and analyzing its performance in cooperation with corresponding institutions
4. Developing a strategy to refer to conditions of traffic incidents, special events, construction works on the roads
5. Developing a methodology for determining the benchmark strategy to be used for comparison to the newly created control strategies.

References

[1]. Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A., & Wang, Y. (2003). Review of road traffic control strategies. *Proceedings of the IEEE*,*91*(12), 2043-2067.

[2]. Abdulhai, B., Pringle, R., & Karakoulas, G. J. (2003). Reinforcement learning for true adaptive traffic signal control. *Journal of Transportation Engineering*, *129*(3), 278-285.

[3]. Abdulhai, B., & Pringle, P. (2003). Autonomous multiagent reinforcement learning-5gc urban traffic control. In *Annual Transportation Research Board Meeting*.

[4]. Koltovska, D. (2013). Q-learning Based Development of Adaptive Signal Control on Isolated Intersections. PhD Thesis, St. Kliment Ohridski University, Bitola Faculty of Technical Sciences Department for Traffic and Transport, Bitola.

[5]. Gartner, N., Stamatindius, C., Tarnoff, P. (1996). Development of Advanced Traffic Signal Control Strategies for ITS. *Transportation Research Record 1494*, 98 -105.

[6]. Koltovska, D., Bombol, K. (2013). Methodology Approach for Developing Adaptive Traffic Control Strategy – A Novel Concept for Traffic Engineers. ISEP 2013, Ljubljana, Slovenia.

[7]. Shoufeng, L., Ximin, L., & Shiqiang, D. (2008, April). Q-Learning for adaptive traffic signal control based on delay minimization strategy. In *Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference on* (pp. 687-691). IEEE.

[8]. Sutton, R.S., Barto, A.G. (1998). *Reinforcement Learning - An Introduction.* MIT Press, Cambridge, Massachusetts.

[9]. Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Doctoral dissertation, University of Cambridge).

[10]. Guberinic, S., Senborn, G., & Lazic, B. (2007). *Optimal traffic control: urban intersections*. CRC Press.

Corresponding author: Daniela Koltovska
Institution: University “St. Kliment Ohridski”, Faculty of Technical Sciences, Department for Traffic and Transport
E-mail: daniela.koltovska@tfb.uklo.edu.mk