



Flávia dos Reis, A., Brante, G., Parisotto, R., Souza, R. D., Valente Klaine, P. H., Battistella, J. P. and Imran, M. A. (2020) Energy efficiency analysis of drone small cells positioning based on reinforcement learning. *Internet Technology Letters*, 3(5), e166. (doi: [10.1002/itl2.166](https://doi.org/10.1002/itl2.166)).

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

This is the peer reviewed version of the following article:

Flávia dos Reis, A., Brante, G., Parisotto, R., Souza, R. D., Valente Klaine, P. H., Battistella, J. P. and Imran, M. A. (2020) Energy efficiency analysis of drone small cells positioning based on reinforcement learning. *Internet Technology Letters*, 3(5), e166, which has been published in final form at [10.1002/itl2.166](https://doi.org/10.1002/itl2.166). This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#).

<http://eprints.gla.ac.uk/214622/>

Deposited on: 24 April 2020

RESEARCH LETTER

Energy Efficiency Analysis of Drone Small Cells Positioning Based on Reinforcement Learning

Ana Flávia dos Reis*¹ | Glauber Brante¹ | Rafaela Parisotto² | Richard Demo Souza² | Paulo Henrique Valente Klaine³ | João Pedro Battistella³ | Muhammad Ali Imran³

¹Federal University of Technology - Paraná (UTFPR), Curitiba, Brazil

²Federal University of Santa Catarina (UFSC), Florianópolis, Brazil

³University of Glasgow, Glasgow, UK

Correspondence

*Ana Flávia dos Reis. Email: anareis@alunos.utfpr.edu.br

Present Address

Federal University of Technology - Paraná (UTFPR). Av. Sete de Setembro, 3165, 80230-901, Curitiba-PR, Brazil.

Summary

This work proposes an algorithm to optimize the positioning and the transmit power of Drone Small Cells (DSCs) based on Q -learning, a reinforcement learning technique where the agents learn to maximize a given reward. We consider two different rewards in this work, the first focusing on maximizing the network coverage, while the second maximizes the lifetime. Then, the Q -learning solution determines the best positioning of the DSC in the 3D space, as well as the optimal transmit power. Results show that the optimization of the transmit power is of paramount importance to reduce the outage probability. In addition, we show that the second reward can considerably increase the network lifetime with a small penalty to the coverage.

KEYWORDS:

Drone Small Cells, Energy Efficiency, Reinforcement Learning

1 | INTRODUCTION

Base Stations (BSs) composed by Drone Small Cells (DSCs) are seen as important components of the new generations of wireless networks in order to provide coverage in a fast and reconfigurable way. DSCs are aerial devices equipped to provide support to a wireless network as mobile BSs. One important aspect of DSC-based systems is the trajectory planning. In this context, the authors in¹ applied convex optimization techniques in order to solve trajectory optimization problems, while² proposed an energy efficient DSC positioning based on the maximization of the network coverage. One important remark is that most of the literature is based on fixed-wing DSCs, which require constant movement. Another model, considering rotary-wing DSC for energy-efficient communication is considered by³, where several advantages are shown when compared to fixed-wing DSCs, such as the ability to take off and land vertically, or to hover, making it more popular and attractive to this market.

Due to the dynamic nature, traffic density and diverse requirements in modern wireless networks, the application of analytical solutions might be infeasible. More flexible solutions that explore the data generated by the network and make decisions in real-time are important for practical deployments. As such, machine learning is a powerful tool to enable self-organization of the communication network using DSCs. For instance, the work in⁴ applied a Q -learning solution in order to optimize the position of DSCs in an emergency communication network scenario. The main goal was to maximize the number of users covered by the communication network, so that a given number of DSCs move to find their best positioning in order to maximize the network coverage. When compared to different positioning strategies, such as positioning the DSCs randomly, around a circle in the center of the scenario, or in the hot spots of the previously destroyed network, the Q -learning solution presented increased performance, minimizing the number of users in outage and converging more quickly. However, energy efficiency is not investigated in⁴, although it is a crucial issue in view of the limited amount of energy and the need for constant recharge of the DSCs.

Another Reinforcement Learning (RL) solution has been used in⁵ to track the DSCs movements in order to maximize the downlink data transmission. As a result, the proposed algorithm provides improved network capacity when compared to the DSC deployment in random static positions. Deep RL was applied in⁶, which was the first study to use such a technique in order to manage interference in a network based on DSCs. The proposed solution reduced up to 62% of the communication latency per DSC, and increased up to 14% in the energy efficiency when compared to a heuristic solution (relying on the shortest path towards the destination). In addition, the minimization of energy consumption for DSC-based networks has also been considered by⁷, in which the goal is to design the optimal rotary-wing DSCs trajectories. Then, alternating optimization and successive convex approximation techniques have been combined in order to minimize the energy consumption.

Different from the prior studies, in this paper we tackle both the outage performance and the energy efficiency of rotary-wing DSCs positioning, so that we employ the power consumption model proposed by³ in a pop-up network scenario, *i.e.*, in which the DSCs provide service for a large concentration of users in a relatively small area, such as such as fairs, musical or sports events. Our solution is based on the Q -learning technique, where the reward can be adjusted to balance between outage and energy goals. Thus, differently from⁷, our approach is not to design optimal trajectories for the DSC, but rather to adapt the positioning of the DSCs autonomously in order to provide both coverage and energy efficiency. In addition, differently from⁴, we also adapt the transmit power of the DSCs in order to reduce the interference, extending previous results from⁸. We show that imposing energy consumption constraints implies in less movements by the DSCs, maximizing the energy efficiency. Also, gains in the autonomy of the DSCs is shown through a network lifetime analysis.

2 | SYSTEM MODEL

System Model: We consider a temporary event scenario, where the number of users to be served is unknown and much higher when compared to the usual deployment of the network in that region. Therefore, in a worst case assumption, we consider that connectivity is provided only by the DSCs. Moreover, the considered urban scenario follows the distribution model of buildings and users defined by the International Telecommunication Union - Radio (ITU-R). According to⁹, the scenario can be described by the ratio of constructed land area to the total area (α), the average number of buildings per km² (β), and a scale parameter for the heights of the buildings (γ) following a Rayleigh probability density function.

For a urban scenario it is assumed that the width of the buildings is given by¹⁰ $W = 1000 \cdot \sqrt{\alpha/\beta}$, while the space between buildings is $S = 1000/\beta - W$, both given in meters. Furthermore, we consider K users randomly distributed in a squared $L \times L$ area, where $\mathcal{K} = \{1, 2, \dots, K\}$ denotes the set of active users.

Communication Model: We consider a set of $\mathcal{D} = \{1, 2, \dots, D\}$ DSCs, in which D is the number of employed DSCs working as aerial BSs. Moreover, each DSC is a single-antenna device, whose aperture angle is denoted by θ . Thus, each DSC $j \in \mathcal{D}$ covers an area given by $\varphi_j = \pi \left(h_j \cdot \tan \frac{\theta}{2} \right)^2$, where h_j is the height of the j -th DSC. The path loss, in dB, between the DSC $j \in \mathcal{D}$ and user $i \in \mathcal{K}$ is given by

$$\kappa_{ij} = 20 \cdot \log_{10} \left(\frac{4\pi f_c d_{ij}}{c} \right) + \varepsilon, \quad (1)$$

where d_{ij} is the distance between the DSC and the user, f_c is the carrier frequency, c is the speed of light in vacuum and ε represents additional losses, which assumes different values depending on the existence or not of line-of-sight (LOS) between the user and the DSC¹¹. Let us remark that the calculation of the presence of LOS or not depends on the instantaneous realization of the urban model, which is done by the simulation framework by taking into account the 3D position of the DSC, the positioning and height of the buildings, as well as the positioning of the users.

We assume all DSCs communicate using the same frequency, causing interference between them. In this sense, the signal to interference plus noise ratio (SINR) for a user i with respect to a DSC j is

$$\rho_{ij} = \frac{P_{r_{ij}}}{B N_0 + \sum_{k=1, k \neq j}^D P_{r_{ik}}}, \quad (2)$$

where B is the system bandwidth, in Hz, N_0 is the noise power spectral density, in W/Hz, and $P_{r_{ij}} = P_t - \kappa_{ij}$ is the reference signal received power, *i.e.*, the power received by the user i when the DSC j employs the transmission power P_t . Moreover, the sum in the denominator of (2) represents the interference caused by neighbor DSCs transmitting at the same time and in the same frequency, whose coverage areas overlap. In addition, the transmit power of each DSC can be adapted according to a few power levels, which is exploited by the algorithm proposed in Section 3.

The association of users with each DSC is made according to their SINRs and resource blocks available, assuming that each user consumes a single resource block of the DSC. If a given user i has SINR ρ_{ij} with respect to the DSC j above a given SINR threshold and the DSC still has available resource blocks to allocate to this user, then that user is associated to that DSC. However, if the DSC in question has no available resource blocks, or if the user SINR is below the threshold, then the next DSC is considered as an alternative to associate that user. Finally, if all DSCs were tested and the user is unable to connect with any of them, then that user is considered in outage for that time slot, *i.e.*, outside the network coverage.

Power Consumption Model: The power consumption related to the movement of the rotary-wing DSCs, in scenarios where the maneuvers have small time duration, depends mainly on the speed of flight V , so that³

$$\xi(V) = \xi_0 \left(1 + \frac{3V^2}{\Omega^2 R^2} \right) + \xi_i \left(\sqrt{1 + \frac{V^4}{4v_0^4}} - \frac{V^2}{2v_0^2} \right)^{\frac{1}{2}} + \frac{1}{2} d_0 \rho s A V^3, \quad (3)$$

where Ω is the blade angular velocity in radians/second, R is the rotor radius in meters, v_0 is the mean rotor induced velocity at the hover, d_0 is the fuselage drag ratio, ρ is the air density, s is the rotor solidity and A is the rotor disc area in m^2 . Besides that, $\xi_0 = \frac{\delta}{8} \rho s A \Omega^3 R^3$ is the power of blade profile, δ is the profile drag coefficient, and $\xi_i = (1 + \zeta) \frac{\chi^{\frac{3}{2}}}{\sqrt{2\rho A}}$ is the induced power, where ζ and χ are the correction factor to induced power and the DSC weight, respectively. Let us remark that the transmit power has not been considered, since the power used to move the DSCs is considerably higher and dominate the analysis³.

3 | PROPOSED ALGORITHM

In this paper we propose a Q -learning based algorithm in order to find the best positioning of DSCs autonomously, as well as their optimal power allocation. In the proposed method we have *a) Agents:* each DSC is an independent agent of the Q -learning solution that is able to move around in an environment composed by building and users; *b) Users:* mobile users, trying to connect with one of the DSCs; *c) States:* a set of possible states for the DSC consists of the 3D positions and the transmission power levels to be used; *d) Actions:* each DSC can take one out of nine possible actions per episode: move forward, backward, up, down, left, right, increase or decrease transmission power, or do nothing.

The goal of the algorithm is to fill a Q -matrix for each independent agent composed by elements $Q(s_t, a_t)$ that represent the value of being in a specific *state* s_t , in the time instant t , while performing a specific *action* a_t . The Q -matrix is updated as

$$Q(s_t, a_t) = Q(s_t, a_t) + \lambda \left[r_{k_{t+1}} + \phi \max_a \{ Q(s_{t+1}, a) \} - Q(s_t, a_t) \right], \quad (4)$$

where λ is the learning rate, $r_{k_{t+1}}$ is the expected reward for the next instant, with $k \in \{1, 2\}$ to denote both proposed rewards, ϕ is the discount factor and $\max_a \{ Q(s_{t+1}, a) \}$ is an estimate function for the optimal value for the future action instant.

Then, we have the following definitions. The **Initialization** of the algorithm defines that the DSCs are randomly positioned and each individual Q -matrix is set to zero. The adopted **Policy** is ϵ -greedy, where ϵ defines the probability that the agent should *explore*, where random actions are taken in different states, and not *exploit*, where the agent uses the acquired knowledge to seek the best possible action within the current state. In addition, the proposed optimization is divided into **Episodes** and each episode is further divided into **Iterations**, which are an instantaneous realization of the environment where the users are approximately static. Thus, the DSCs are able to test some actions and observe the obtained reward. Finally, we modify the Q -learning solution in order to provide **Stopping Criteria**, so that the actions are carried out until one of the following stop criteria are met: *(i.)* the number of users covered by the wireless network has not improved in a certain number of iterations ($t_{\text{out,max}}$); *(ii.)* the DSC has moved for a maximum number of iterations (t_{max}); *(iii.)* the DSC has used all its resource blocks. When the DSC meets one of these conditions, it moves to the state that results in the highest reward, so that all DSCs are optimized in a sequence, until the next episode begins. In addition, when the next episode begins, the DSCs start in the last position of the previous episode.

The proposed solution is summarized in Algorithm 1. Moreover, the metrics considered in this paper are the percentage of users in outage and the DSCs energy efficiency. The percentage of users in outage is given by

$$U_{\text{out}} = 1 - \frac{\sum_{j=1}^D U_j}{K}, \quad (5)$$

where U_j is the number of users allocated to the DSC j . In addition, the energy efficiency of the DSCs is defined by

$$\eta = \frac{R_b \cdot (1 - U_{\text{out}})}{E_{\text{total}}}, \quad (6)$$

Algorithm 1 Proposed Q -learning algorithm.

-
- 1: Initialize DSCs in random positions and the Q -matrix with null elements
 - 2: **for every episode do**
 - 3: **while step criteria are not met do**
 - 4: DSCs select the largest $Q(s_t, a_t)$, moving to the respective state s_t performing action a_t
 - 5: DSCs allocate users and observe $r_{k_{t+1}}$
 - 6: Each Q -matrix is updated according to (4)
 - 7: **end while**
 - 8: **end for**
-

given in [bps/J] once R_b is the bit rate, in [bps], and E_{total} is the total energy spent to perform all the DSCs movements, considering the sum of all iterations, *i.e.*, all instants when network users are considered to be static during the execution of the algorithm. Denoting \mathcal{I} as the set of performed iterations, the total power consumption is given by

$$P_{\text{total}} = \sum_{l \in \mathcal{I}} \sum_{j \in \mathcal{D}} \xi_{j,l} \quad (7)$$

where $\xi_{j,l}$ is the power consumed to move the DSC j in the current iteration l , calculated using (3). Therefore, the total energy consumption is given by $E_{\text{total}} = P_{\text{total}} \cdot \Delta_t$, where $\Delta_t = \frac{d_{j,l}}{V}$ is the time elapsed to perform this movement, with $d_{j,l}$ being the distance traveled by the DSC in the current iteration with speed V .

3.1 | Rewards

Furthermore, in order to obtain the maximization of users covered by the network, a first proposed reward is defined as

$$r_1 = \sum_{j=1}^D U_j, \quad (8)$$

recalling that U_j is the number of users connected to each DSC j . This reward regards only the number of users covered by the network and, therefore, leads to an algorithm that aims to minimize U_{out} in (5). In addition, let us remark that r_1 is the same for all DSCs and it is computed locally, considering a backhaul connection among the DSCs.

A second proposed reward is a generalization of r_1 , so that we define

$$r_2 = \left(\sum_{j=1}^D U_j \right) - w \cdot (\xi_{j,l} \Delta_t), \quad (9)$$

which is also computed locally, but is independent from one DSC to another. In addition, w is the weight given to the energy in the reward, which depends on the number of users (K) and may be adjusted so that $\sum_{j=1}^D U_j$ and $\xi_{j,l} \Delta_t$ have similar orders of magnitude. Since this reward presents a trade off between the network coverage and the energy consumed to move the DSCs, the proper selection of w is important in order to minimize U_{out} with an adequate penalty for each DSC movement.

4 | SIMULATION RESULTS

The simulation scenario consists of an urban area of $L = 500$ [m] and $K = 200$ users. In each round of the simulation the users positions are randomly initiated and 100 rounds of 100 episodes are executed in an independent way. Each DSC can serve up to 50 users and the step movement for the X and Y axis is of 50 [m], while 100 [m] is considered for the Z axis. Moreover, h_j is delimited in the range of 100-1000 [m]. ITU-T parameters follow⁹, with $\alpha = 0.3$, $\beta = 500$ [buildings/km²] and $\gamma = 15$ [m]. Additional losses are considered as¹⁰ $\varepsilon = 1$ [dB] for LOS and $\varepsilon = 20$ [dB] for NLOS. In addition, $\theta = 60^\circ$, $B = 180$ [kHz], $f_c = 1$ [GHz], and $R_b = 360$ [kbps]. For the Q -learning algorithm we consider a learning rate of $\lambda = 0.9$, discount factor $\phi = 0.9$ and exploit/explore ratio of $\epsilon = 0.5$. Moreover, the stopping criteria are $t_{\text{max}} = 3600$ and $t_{\text{out,max}} = 20$. Finally, with respect to the power consumption model we follow³, so that the DSC speed is $V = 5$ [m/s], $\chi = 20$ [N], $\rho = 1.225$ [kg/m³], $R = 0.4$ [m], $A = 0.503$ [m²], $\Omega = 300$ [rad/s], $v_0 = 4.03$ [m/s], $d_0 = 0.6$, $s = 0.05$, $\delta = 0.012$ and $\zeta = 0.1$.

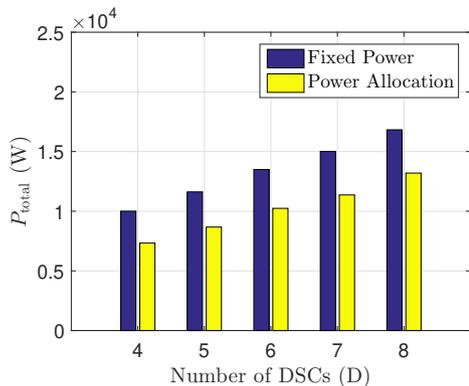


FIGURE 1 Power consumed to move the DSCs.

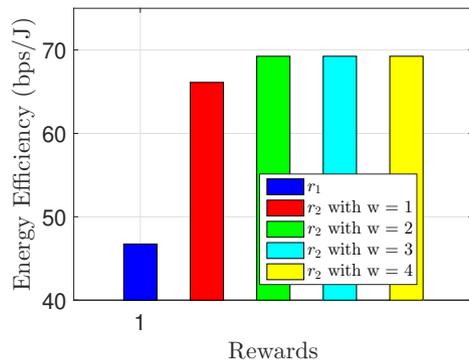


FIGURE 2 Energy efficiency for different w .

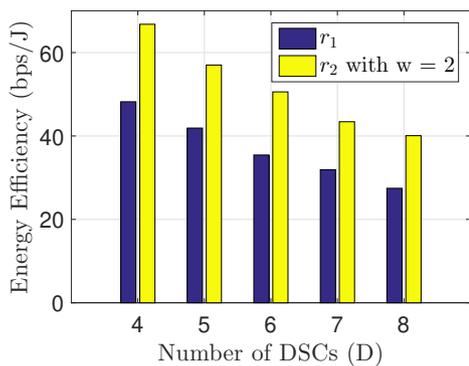


FIGURE 3 Energy efficiency for r_1 and r_2 .

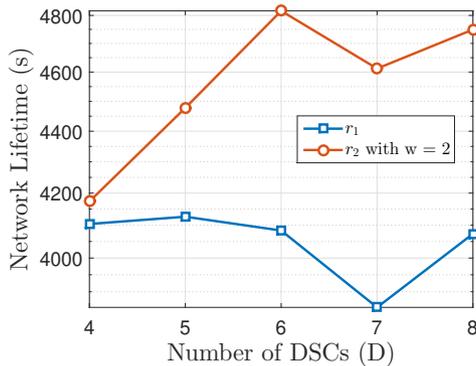


FIGURE 4 Network lifetime vs. D .

Extending⁸, we consider that the DSCs can choose three possibilities for the transmission power: -20 [dB], -10 [dB] and 0 [dB], what can improve the number of allocated users to the DSCs, due to the lower interference among the DSCs when compared to the case of fixed transmit power. Moreover, it is also shown in⁸ that both solutions quickly converge after around 10 episodes. Therefore, in order to demonstrate only the steady-state performance of the proposed algorithm, a similar strategy is adopted here, and the numerical results presented consider an average starting at the 11th episode.

Figure 1 shows the total power consumption to move the DSCs considering reward r_1 . As we can see, the power allocation for the DSCs results in lower average power consumption. In particular, the fixed power algorithm has to change the height of the DSCs in order to reduce interference between them⁴. On the other hand, the proposed algorithm is able to allocate different power levels with the same goal, with the advantage of reducing the energy consumption to move the DSCs at the same time. In addition, Figure 2 shows the energy efficiency in a scenario with $D = 4$ DSCs, comparing r_1 and r_2 with $w \in \{1, 1.5, 2, 2.5, 3\}$. As we observe, $w = 2$ is the weight that yields the best result in terms of energy efficiency. For $w > 2$, η saturates at the same value, indicating that the DSCs are already traveling the shortest possible distances in these situations. Similar conclusions are obtained with different number of DSCs and, thus, the following results consider r_2 with $w = 2$.

Figure 3 investigates the energy efficiency. As we can observe, the reward r_2 yields higher energy efficiency than r_1 regardless of D . For instance, with $D = 8$ DSCs we observe an increase of 36.91% in terms of energy efficiency comparing r_2 with r_1 . This result highlights the importance of the parameter w in the reward of the Q -learning solution in terms of energy efficiency. On the other hand, it is expected that the percentage of users in outage increases. In the same scenario, comparing r_2 with $w = 2$ and r_1 , U_{out} increases 12.69% when $D = 7$ DSCs are employed. Nevertheless, the percentage of users in outage considering $D = 7$ DSCs with the proposed Q -learning algorithm using r_2 and $w = 2$ is only 0.14%, which justifies the savings in terms of energy consumption. Notice that this is different from the conclusions in^{4,8}, since the optimal D to maximize η tends to be as small as possible, while D increases when the goal is to minimize U_{out} in order to balance user coverage and interference among DSCs. As a final analysis, we consider the network lifetime by assuming the battery consumption of the DSCs. Following¹², the initial

battery energy of each DSC is considered to be of 100 [Wh]. Then, the energy consumption for each movement of the DSCs is decreased from this initial value, so that we establish a stop criterion when the first DSC reaches 10% of battery charge. In this situation, this DSC returns to base for recharging its battery and the network lifetime is computed at this time instant. Figure 4 illustrates this analysis considering the rewards r_1 and r_2 with $w = 2$. As we observe, r_2 increases the network lifetime up to 19.66% when $D = 7$ DSCs are employed, which is a considerable gain in terms of flight autonomy.

5 | CONCLUSION

This paper proposes a Q -learning solution for the optimal positioning of DSCs. Our proposal is a multi-objective approach, once both number of users in outage and energy consumption due to the movement of the DSCs are taken into account. Results show that constraining the number of movements of the DSCs is important to maximize the energy efficiency of the network. Moreover, our network analysis has shown that up to 19.66% of increase in the network lifetime can be obtained, which is crucial for the network deployment in temporary events. As future works, we plan to extend the network users requirements, considering restrictions related to speed and latency, as well as to consider different RL techniques to be compared with the Q -learning.

References

1. Zeng Y, Zhang R. Energy-Efficient UAV Communication With Trajectory Optimization. *IEEE Transactions on Wireless Communications* 2017; 16(6): 3747-3760.
2. Alzenad M, El-Keyi A, Lagum F, Yanikomeroğlu H. 3-D Placement of an Unmanned Aerial Vehicle Base Station (UAV-BS) for Energy-Efficient Maximal Coverage. *IEEE Wireless Communications Letters* 2017; 6(4): 434-437.
3. Zeng Y, Xu J, Zhang R. Energy Minimization for Wireless Communication With Rotary-Wing UAV. *IEEE Transactions on Wireless Communications* 2019; 18(4): 2329-2345.
4. Klaine PV, Nadas JP, Souza RD, Imran MA. Distributed Drone Base Station Positioning for Emergency Cellular Networks Using Reinforcement Learning. *Cognitive Computation* 2018; 10(5): 790–804.
5. Wang Q, Zhang W, Liu Y, Liu Y. Multi-UAV Dynamic Wireless Networking With Deep Reinforcement Learning. *IEEE Communications Letters* 2019; 23(12): 2243-2246.
6. Challita U, Saad W, Bettstetter C. Interference Management for Cellular-Connected UAVs: A Deep Reinforcement Learning Approach. *IEEE Transactions on Wireless Communications* 2019; 18(4): 2125-2140.
7. Zhan C, Lai H. Energy Minimization in Internet-of-Things System Based on Rotary-Wing UAV. *IEEE Wireless Communications Letters* 2019; 8(5): 1341-1344.
8. Parisotto RP, Klaine PV, Nadas JPB, Souza RD, Brante G, Imran MA. Drone Base Station Positioning and Power Allocation using Reinforcement Learning. In: International Symposium on Wireless Communication Systems; 2019: 213-217.
9. ITU-R . Propagation Data and Prediction Methods for The Design of Terrestrial Broadband Millimetric Radio Access Systems. 2003: 1410–2.
10. Al-Hourani A, Kandeepan S, Jamalipour A. Modeling air-to-ground path loss for low altitude platforms in urban environments. In: IEEE Global Communications Conference; 2014: 2898-2904.
11. Mozaffari M, Saad W, Bennis M, Debbah M. Unmanned Aerial Vehicle With Underlaid Device-to-Device Communications: Performance and Tradeoffs. *IEEE Transactions on Wireless Communications* 2016; 15(6): 3949-3963.
12. Galkin B, Kibilda J, DaSilva LA. UAVs as Mobile Infrastructure: Addressing Battery Lifetime. *IEEE Communications Magazine* 2019; 57(6): 132-137.

